Behav Res (2017) 49:513–537
DOI 10.3758/s13428-016-0740-2

How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria

Veronika Lerche[1] · **Andreas Voss**[1] · **Markus Nagler**[1]

Published online: 10 June 2016

Abstract Diffusion models (Ratcliff, 1978) make it possible to identify and separate different cognitive processes underlying responses in binary decision tasks (e.g., the speed of information accumulation vs. the degree of response conservatism). This becomes possible because of the high degree of information utilization involved. Not only mean response times or error rates are used for the parameter estimation, but also the response time distributions of both correct and error responses. In a series of simulation studies, the efficiency and robustness of parameter recovery were compared for models differing in complexity (i.e., in numbers of free parameters) and trial numbers (ranging from 24 to 5,000) using three different optimization criteria (maximum likelihood, Kolmogorov–Smirnov, and chi-square) that are all implemented in the latest version of fast-dm (Voss, Voss, & Lerche, 2015). The results revealed that maximum likelihood is superior for uncontaminated data, but in the presence of fast contaminants, Kolmogorov–Smirnov outperforms the other two methods. For most conditions, chi-square-based parameter estimations lead to less precise results than the other optimization criteria. The performance of the fast-dm methods was compared to the EZ approach (Wagenmakers, van der Maas, & Grasman, 2007) and to a Bayesian implementation (Wiecki, Sofer, & Frank, 2013). Recommendations for trial numbers are derived from the results for models of different complexities. Interestingly, under certain conditions even small numbers of trials ($N < 100$) are sufficient for robust parameter estimation.

Keywords Diffusion model · Fast-dm · Mathematical models · Reaction time methods

The diffusion model was introduced almost four decades ago by Roger Ratcliff (1978) as a model for cognitive processes in memory retrieval. Since then, it has been shown that the model can map cognitive processes from a multitude of different cognitive tasks that require fast binary decisions, including—for example—color or numerosity classifications, or lexical decision tasks (see Voss, Nagler, & Lerche, 2013, for a recent review). Thus, the diffusion model can be seen as a generic model for binary decisions. Why have accumulator models like the diffusion model become so popular in recent years? The advantage over traditional analyses of response time (RT) means (or error rates) is that different aspects of cognitive processing can be measured separately. Imagine a study on cognitive aging that analyzes the stability (or decline) of cognitive performance in a specific task at high age. If the mean RT is used as the dependent measure, you cannot be sure whether the longer RTs are really based on slower information processing, because older adults may be more cautious—that is, they may respond only if they are really sure about the correct response (e.g., Forstmann et al., 2011; Ratcliff, Thapar, Gomez, & McKoon, 2004). Additionally, older adults may be slower in motor response execution (e.g., Ratcliff, Thapar, & McKoon, 2004). Thus, it is important to get a valid measure for speed of information processing that is not confounded by speed–accuracy settings or the speed of motor response. In the diffusion model framework, the drift parameter provides such a measure of cognitive speed (see Voss,

✉ Veronika Lerche
veronika.lerche@psychologie.uni-heidelberg.de

[1] Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Hauptstrasse 47-51, D-69117 Heidelberg, Germany

Rothermund, & Voss, 2004, for an experimental validation study).

In the first three decades following its introduction in psychological research, Ratcliff's (1978) diffusion model was used primarily by researchers with a profound interest in, and knowledge of, mathematical psychology. In recent years, however, the diffusion model has increasingly attracted the attention of researchers from various other fields of psychology. Examples indicating the wide range of applications for the diffusion model include analyses of cognitive processes in such typical experimental paradigms as the lexical decision task (e.g., Yap, Balota, & Tan, 2013), sequential priming paradigms (e.g., Voss, Rothermund, Gast, & Wentura, 2013), task switching (Schmitz & Voss, 2012, 2014), or prospective memory paradigms (e.g., Boywitt & Rummel, 2012). Other applications encompass social cognitive research (e.g., Germar, Schlemmer, Krug, Voss, & Mojzisch, 2014; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Voss, Rothermund, & Brandtstädter, 2008), cognitive aging (e.g., McKoon & Ratcliff, 2013; Spaniol, Madden, & Voss, 2006), cognitive processes related to psychological disorders (e.g., Metin et al., 2013; Pe, Vandekerckhove, & Kuppens, 2013; White, Ratcliff, Vasey, & McKoon, 2010b), and other fields of psychology.

So far, the diffusion model has often been applied for the detection of differences between groups or conditions (e.g., Boywitt & Rummel, 2012). More recently, the correlations between diffusion model parameters and external criteria have also constituted a research field (e.g., between drift rate and general intelligence; see Ratcliff, Thapar, & McKoon, 2010). On the basis of such observations, lately, the idea has been expressed that the diffusion model might also be used as a diagnostic tool (e.g., Aschenbrenner, Balota, Gordon, Ratcliff, & Morris, 2016; Ratcliff & Childers, 2015).

These different types of applications of the diffusion model go along with different requirements regarding parameter estimation accuracy. For example, for the detection of differences between conditions, biases in parameter estimation are not necessarily a problem. Imagine an estimation procedure that results in a systematic overestimation of the drift rates in both of two conditions. If the estimation bias is similar over conditions, it will not affect the power of difference detection. If, however, the estimation bias depends on the experimental condition (e.g., via the number of error responses), the power to detect differences between the conditions might be affected. In another scenario, there might be no systematic estimation bias, but imprecise measurement could lead to large average deviations between the true and reestimated parameter values. The increased error variance would directly diminish the power of difference detection. In this case, an increase in the number of participants can reestablish the power to detect any effects on parameters. Finally, if the diffusion model is applied for the diagnosis of interindividual differences in cognitive functioning, it is important that the relevant parameter be estimated very accurately (i.e., reliably) for each single individual. Thus, depending on the aim of the researcher, more or less strict criteria would have to be applied.

One important methodological factor that directly influences the precision of results is the number of trials. There has been a huge variation in the numbers of trials used for previous diffusion model experiments, ranging from less than 100 to several thousands of trials per participant. The choice of trial numbers typically seems to be rather arbitrary. It is remarkable that the required trial numbers have rarely been analyzed systematically so far (see Lerche & Voss, in press; Ratcliff & Childers, 2015; and Wiecki, Sofer, & Frank, 2013, for some exceptions).

The main aim of the present article is to provide well-founded recommendations regarding the requisite trial numbers for robust diffusion modeling. As we discussed above, the question of requisite trial numbers is closely related to the precision that is necessary for a specific research question. In a series of simulation studies, we tested the precision of parameter estimation procedures for very small to very high trials numbers. This allowed us to derive conclusions for minimally required trial numbers (i.e., a number below which diffusion modeling becomes virtually meaningless), as well as "maximum" trial numbers (above which increases in precision become negligible).

A factor influencing the required number of trials is the efficiency of the applied estimation procedure. Accordingly, a second objective of this article is the comparison of the efficiency of different optimization criteria for the parameter search procedure for diffusion modeling. The simulations in this article were carried out using *fast-dm-30* (Voss, Voss, & Lerche, 2015), which is the newest version of fast-dm (Voss & Voss, 2007, 2008). Besides the Kolmogorov–Smirnov criterion that was implemented in former versions, fast-dm-30 now includes implementations of the chi-square and maximum likelihood criteria. The implementation of these within the same program facilitates comparisons of the criteria's performance. Thus, in contrast to studies that have compared different programs (e.g., van Ravenzwaaij & Oberauer, 2009), we can exclude the possibility that any differences between optimization criteria are due to program specifics.

Finally, a third focus is the influence of model complexity on the required numbers of trials. Typically, diffusion model analyses allow for intertrial variations of the diffusion model parameters (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). Although estimates of intertrial variability seldom allow for meaningful psychological interpretations, they often do improve model fit. However, it remains unclear how this increase in model complexity would influence the precision of estimates for the more meaningful diffusion model parameters. To investigate the influence of model complexity, we analyzed four differently complex models. Note that in the

present article only models for simple experimental designs are considered. If data from more complex designs with different conditions were mapped, models would probably be more stable (and hence, the requisite trial numbers lower) if it were known on which parameters the manipulation would map; if not, the increasing number of model parameters might make the model even more unstable.

In the following sections, we first give a short introduction to diffusion modeling. This is followed by the presentation of the main properties of the different optimization criteria (i.e., chi-square, maximum likelihood, and Kolmogorov–Smirnov). In the subsequent section, the available computer programs for diffusion model analyses are briefly presented. After this, we go into the main research issues, giving an overview of initial simulation studies comparing different optimization criteria and different trial numbers. Finally, we outline and discuss the methods and results of our simulation studies.

The rationale of the diffusion model

Researchers dealing with data from binary decision tasks often use either the percentage of correct responses or the mean RTs as dependent measures. However, some research questions cannot be properly addressed on the sole basis of (one of) these measures. For instance, different speed–accuracy settings can make it difficult to interpret an observed difference in mean RTs between two groups or conditions. Are the longer RTs in one of these conditions due to slower information uptake, or rather the result of a conservative response style? The diffusion model helps solve this problem, because it maps speed–accuracy settings and the speed of information processing on independent parameters. This decomposition becomes possible by taking into account the complete distributions of both correct and error responses (and thus, implicitly, also the error rate). Thereby, several cognitive components are identified that have clear psychological interpretations (Voss et al., 2008; Voss et al., 2004). This makes it possible to answer not only the question of *whether or not* people (or tasks) differ in their performance in a cognitive task, but also to determine *in what way* they differ (e.g., *why* one person is faster than another). Note that several mathematical models allow such a separation of the different components involved in decision tasks. One prominent example is the linear ballistic accumulator model (Brown & Heathcote, 2008). In this article, we focus on the diffusion model (Ratcliff, 1978).

The basic assumption of the diffusion model is that decisions are based on a continuous information-sampling process that is described by a Wiener diffusion process (i.e., a diffusion process with constant drift) running in a corridor between two thresholds (see Fig. 1). The current information drives the decision process toward the upper or the lower threshold,

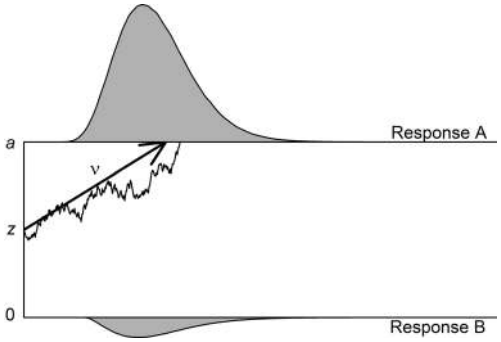

Fig. 1 Example illustration of the decision process of the diffusion model. The process starts at z (here situated in the middle of threshold distance a) and moves with mean drift rate v until a threshold is hit (here the upper threshold). In the following, the motoric execution of the associated response (here Response A) is initiated

representing two possible decisional outcomes. As soon as the upper or lower threshold is hit, the decision is reached, and a corresponding motor program is initiated. Because the diffusion process is a stochastic (i.e., noisy) process, durations and outcomes may vary from trial to trial, even if identical stimulus information is presented.

In the following paragraphs, we briefly present the parameters of the diffusion model. The *drift rate* (parameter ν) indicates the average speed (and direction) of information uptake. High (absolute) drift rates lead to fast responses and few errors, whereas a drift around zero indicates chance performance with long RTs. Thus, high drift rates indicate higher cognitive speed or easy tasks.

A second model parameter is the *distance between the two thresholds* (parameter a). This parameter defines how much information is considered before a decision is made. A large threshold separation means that a lot of information needs to be sampled before the decision is made, which will result in large RTs with a low error rate. Thus, conservative decision-makers will have large threshold separations, and liberal decision-makers small ones.

The *starting point* (parameter z) defines the position at which information accumulation begins. If z is not centered between the thresholds, there is a decision bias in favor of the threshold that is closer to the starting point. To reach this "preferred" threshold, the process needs less information, and the corresponding responses will therefore be more frequent and faster. Instead of the absolute value z, often the relative starting point $z_r = z/a$ is reported (e.g., Voss et al., 2015), with $z_r = .5$ reflecting an unbiased decision process. Note that the decision bias mapped by the starting point is conceptually similar to response bias in the signal detection framework. We prefer the term *decision* bias because it influences the decision process, and not merely response execution.

In addition to the decision times, in the analysis of RT data the duration of *nondecisional processes* (parameter t_0 or T_{er},

not shown in Fig. 1) also needs to be considered. These nondecisional processes can temporally precede (e.g., encoding of information) or follow (motoric execution of the response) the accumulation process.

Furthermore, the diffusion model can also explain trial-to-trial fluctuations in performance that arise—for example—from variability in the stimulus information or in the attention of the participant. For this purpose, intertrial variability parameters have to be included (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002; see also Laming, 1968). Specifically, it is assumed that the drift across trials follows a normal distribution with mean ν and standard deviation s_ν. The starting point and nondecision time are assumed to be normally distributed, with means z_r and t_0 and widths s_{zr} and s_{t0}, respectively. More recently, the diffusion model has been expanded to include a response bias parameter (parameter d) that maps differences in the duration of nondecisional processes between the two responses (Voss, Voss, & Klauer, 2010; Voss et al., 2015).

Finally, the diffusion model includes the diffusion coefficient—that is, the amount of noise in the diffusion process (sometimes called the *intratrial variability of drift*). The diffusion coefficient is typically not estimated but instead used as a scaling parameter (theoretically, either ν or a could be used as the scaling parameter, and then the diffusion coefficient could be estimated). We set the diffusion coefficient to $s = 1$ (and thus held it constant across conditions; see Donkin, Brown, & Heathcote, 2009, for a different suggestion). If another value is used, all diffusion parameters (except t_0 and s_{t0}) are rescaled by the factor s (e.g., Ratcliff usually sets s to .1 in his applications).

Optimization criteria

One common aim in diffusion model analysis is to find a set of parameters that optimally describes the empirical data. To achieve this, deviations between the observed data and the data predicted from a certain set of parameters are minimized by adjusting the parameter values. For this purpose, different optimization criteria quantifying the goodness of fit between the observed and expected data have been used in the diffusion model literature. In the following discussion, we present three criteria that have frequently been applied in the context of diffusion modeling: chi-square (CS), maximum likelihood (ML), and Kolmogorov–Smirnov (KS) (see also Table 1).

Chi-square The CS criterion has often been used for parameter estimation (Ratcliff & McKoon, 2008; Ratcliff & Tuerlinckx, 2002; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). To calculate CS, responses are grouped into bins according to latency. This is done separately for the responses at the upper and lower thresholds. The borders of the bins are based on quantiles of the RTs observed. Ratcliff and Tuerlinckx (2002) proposed the use of six bins, with the two outer bins each comprising 10 % of the observed RTs, and the other four bins 20 % each. Accordingly, the borders of the RT bins are defined by the 10th, 30th, 50th, 70th, and 90th percentiles of the empirical RT distributions. These bins are then applied to the predicted distributions. From the deviations between the numbers of predicted and observed responses for each bin, a CS value is computed.[1] In an iterative parameter search process, this CS sum is minimized. Because the predicted (cumulative) distributions only need to be evaluated at the borders of the bins, computation is fast and independent of the number of trials.

Maximum likelihood The ML criterion is used in various mathematical modeling approaches. In contrast to the CS approach (e.g., Ratcliff & Tuerlinckx, 2002), the ML approach uses every RT, and no binning is necessary. A set of parameters is sought to maximize the likelihood of the empirical data. For technical reasons (see Ratcliff & Tuerlinckx, 2002), typically the sum of logarithmized density values is maximized rather than the product of densities. Unlike CS, the computation time required by ML depends strongly on the number of trials per data set, because the predicted density has to be computed for each trial.

Kolmogorov–Smirnov The KS criterion was introduced as an optimization criterion for diffusion modeling by Voss et al. (2004) and has been applied in numerous studies (e.g., Horn, Bayen, & Smith, 2011; Metin et al., 2013; Voss, Rothermund, et al., 2013). The criterion is based on the cumulative distribution functions (CDFs) of RTs. To calculate the KS criterion, the distributions at the upper and lower thresholds are combined by multiplying all RTs from the lower threshold by –1 (a procedure first proposed by Voss et al., 2004; see also Voss & Voss, 2007). This creates a cumulative density function for the whole data set. The KS criterion is the maximum absolute vertical distance between the observed and predicted CDFs. Accordingly, for each observed RT the distance between the two CDFs needs to be computed to identify the maximum. The iterative search for parameter estimates then aims at minimizing this maximum distance.

[1] Strictly speaking, the resulting value is not exactly a chi-square value, because the borders of the bins are determined from the empirical distributions and not from the predicted distributions (Speckman & Rouder, 2004). However, the values approximate a chi-square distribution (Fifić, Little, & Nosofsky, 2010). In diffusion model analyses, CS is usually calculated in this way, because it is computationally much easier and faster.

Table 1 Comparison of optimization criteria

	Chi-Square	Maximum Likelihood	Kolmogorov–Smirnov
Term to be minimized in the optimization process	$\sum \frac{(o_i - p_i)^2}{p_i}$ Note: o_i /p_i correspond to the numbers of responses observed/predicted in bin i	$-\sum ln(d(RT_i, k_i))$ Note: $d(RT_i, k_i)$ corresponds to the density value of the RT observed in trial i with response k_i	$\max_{i=1\ldots n} \left\lvert \begin{matrix} eCDF(RT_i) - \\ pCDF(RT_i) \end{matrix} \right\rvert$ Note: n is the number of responses observed; eCDF/pCDF are the empirical/predicted cumulative distribution functions; RT_i is the RT in trial i
Information utilization	low	high	medium
Computation time	low	high	medium

Estimation programs

For many years researchers had to develop parameter search implementations of their own for the diffusion model analyses. In recent years, several programs were published for this purpose. Among them is the *EZ-diffusion model* (Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, van der Maas, Dolan, & Grasman, 2008; Wagenmakers, van der Maas, & Grasman, 2007), which is available as JavaScript, R code, a MATLAB implementation, and an Excel spreadsheet. In comparison with search procedures based on the three optimization criteria presented in the last section, EZ uses a more limited amount of information. In the original version of EZ (Wagenmakers et al., 2007), parameters were estimated from error rates and the mean and variance of the correct responses. Closed-form equations are utilized for the parameter calculation. In this way, estimates for three parameters can be obtained (a, ν, t_0). In the extended versions of EZ (Grasman et al., 2009; Wagenmakers, van der Maas, et al., 2008), further parameter options are available, such as estimation of the parameter z and the consideration of contaminant data.

The *Diffusion Model Analysis Toolbox* (DMAT; Vandekerckhove & Tuerlinckx, 2007, 2008) is a MATLAB toolbox. In DMAT, the CS method is implemented. Furthermore, the toolbox offers the possibility of using quantile maximum probability estimation (see also Heathcote & Brown, 2004; Speckman & Rouder, 2004).

A third program, *fast-dm* (Voss & Voss, 2007, 2008), is a command-line program. In fast-dm-29 and all earlier versions, parameter search was generally based on KS as the optimization criterion. The newest version, *fast-dm-30* (Voss et al., 2015), offers a choice between the KS, ML, and CS approaches.

The last few years have seen the advent of software solutions for hierarchical diffusion model analyses. Vandekerckhove, Tuerlinckx, and Lee (2011) proposed a plug-in to the WinBUGS software. A platform-independent solution, HDDM (for *hierarchical drift diffusion model*) has been presented by Wiecki et al. (2013). HDDM is a toolbox based on Python and uses a Bayesian method for parameter estimation. It can be used either for fitting a hierarchical model or for fitting parameters for each individual subject. Recently, another platform-independent software option was introduced by Wabersich and Vandekerckhove (2014).

Literature on comparison of optimization criteria

There is a lack of systematic research on the performance of different optimization criteria for diffusion modeling. One exception is the study by van Ravenzwaaij and Oberauer (2009), who compared the performance of EZ, fast-dm, and DMAT (using the multinomial log-likelihood function, MLF). They found KS to be superior to MLF in terms of the correlations between the true and recovered parameter values. However, MLF performed better than KS in recovering the mean true values. Since the comparison of the optimization criteria KS and MLF was based on different software solutions, however, program details may have been the factor behind the resulting differences, which were not necessarily based on the different optimization criteria. Interestingly, EZ performed very well in this study. Especially in the event of a reduction in the number of trials (80 instead of 800), EZ outperformed fast-dm in the correlations (DMAT could not even be applied in this condition, since it needs a minimum of 11 errors in each RT quantile). However, EZ (even in the more recent versions) does not allow for the estimation of intertrial variabilities, so full comparability with KS and MLF cannot be established. Note that the estimation of intertrial variabilities (especially s_z and s_ν) posed serious problems for fast-dm and DMAT. This may have had a negative influence on the recovery of the other parameters. EZ, on the other hand, circumvents the estimation difficulties associated with intertrial variabilities by providing estimates for only three parameters (a, ν, and t_0). Besides, no contaminated trials were included in the simulation studies. *Contaminants* are responses resulting from sources other than a diffusion process. In several simulation studies, Ratcliff (2008) demonstrated the sensitivity of EZ to the presence of

contaminants (but see Wagenmakers, van der Maas, et al., 2008).

Ratcliff and Tuerlinckx (2002) compared ML, CS, and a weighted least squares (WLS) fitting method, both with and without the inclusion of contaminants. They showed that for a model consisting of eight parameters (a, t_0, four drift rates, s_z, and s_v; z was assumed to be centered between the thresholds) and data without contaminants, ML outperformed CS and WLS. When contaminants were added, ML's performance deteriorated dramatically. CS was also impaired by the presence of contaminant trials, whereas the performance of WLS deteriorated only slightly. Ratcliff and Tuerlinckx counteracted the deterioration of ML and CS by explicitly modeling the contaminants with a uniform distribution. Consequently, parameter recovery improved. Furthermore, they included the intertrial variability of t_0 into the model (s_{t0}), and with both this additional parameter and the modeling of contaminants, CS resulted in precise and unbiased estimation when 1,000 trials per condition were used. For 250 trials per condition, the performance was significantly worse. The authors recommended using CS with the correction for contaminant trials and s_{t0} included in the model. Note, however, that for 250 trials per condition, the performance of CS in this model was "very poor" (p. 467). Subsequently, many researchers have used CS, referring to the studies by Ratcliff and Tuerlinckx, and stated that CS "provides the best balance between robustness and the ability to recover parameter values" (Wagenmakers, Ratcliff, et al., 2008, p. 146).

Recently, the performance of newly developed hierarchical diffusion models has been tested. Wiecki et al. (2013) compared CS- and ML-based algorithms to HDDM, their software solution for a hierarchical Bayesian estimation of parameters. Their work revealed the superiority of HDDM, especially for small trial numbers. Besides, ML often outperformed CS.

Ratcliff and Childers (2015) ran a series of simulation studies in which they compared eight different estimation methods and programs. DMAT cut a poor figure, and EZ did not perform very well in the presence of contaminants. However, CS (based on either ten or six bins), ML, and KS generally recovered the parameters quite well. Some of the findings for HDDM were inconsistent. For example, in Simulation Study 1, in one design (with four drift rates) HDDM featured high correlations between the true and reestimated parameter values even for small trial numbers, outperforming the other methods. In another design (with two drift rates) for smaller numbers of trials, it performed worse than most of the other methods. Besides, in another simulation study (Simulation Study 2), unexpectedly, high biases were found for a large trial number, whereas the biases for a small trial number were smaller than those in the other methods.

With the availability of fast-dm-30 (Voss et al., 2015), the three criteria CS (based on six bins), ML, and KS can be compared to each other independently of confounding factors (i.e., program specifics). As we outlined in the section on optimization criteria, CS, ML, and KS differ in the amounts of information used for the fitting process. Whereas CS reduces the available information by dividing the distributions into bins, ML and KS consider the exact value of each RT observed. This is why we expected our simulation studies to reveal that the number of trials required for efficient parameter estimation was higher for the CS criterion than for ML and KS. KS requires calculation of the vertical distance for each RT observed; the criterion itself, however, is based on only one of these distances (the maximum distance). Accordingly, ML may be a more efficient estimator than KS.

However, we also expected the optimization criteria to differ in terms of robustness in the presence of "contaminants."[2] Because the (log-)likelihood can be strongly influenced by single RTs, we assumed that results from the ML method would be most strongly biased when RT distributions were contaminated, whereas the CS and KS criteria were expected to be more robust. Therefore we expected ML to require the lowest number of trials, followed by KS and CS, with uncontaminated data. In the presence of contaminants, however, ML should perform worse than KS.

Number of trials required

Is diffusion modeling restricted to experimental designs with more than 1,000 trials per participant? After receiving regular inquiries from researchers greatly interested in diffusion modeling but uncertain about the number of trials required for robust analysis, we decided to address this issue systematically. Conventionally, high numbers of trials are used for diffusion modeling. For instance, Ratcliff, Thapar, Gomez, et al. (2004) used 2,100 trials in their experimental session (see also Leite & Ratcliff, 2011; Ratcliff & Rouder, 1998; Ratcliff & Van Dongen, 2009; Wagenmakers, Ratcliff, et al., 2008; but see Klauer et al., 2007). Although generally a large database makes the fitting of mathematical models more stable, obviously using extraordinarily large trial numbers can cause problems of its own. First, the experimental sessions require more time and effort. More importantly, psychological effects may change over time due to practice effects, and after several hundred trials, some effects of interest may be diminished, or even disappear completely. Additionally, it may often be difficult to find sufficient stimuli, if they are not supposed to be repeated.

One interesting approach to addressing the issue of trial numbers by way of experimental design has been proposed by White, Ratcliff, Vasey, and McKoon (2009), who used filler trials (see also White et al., 2010b) to achieve higher

[2] We consider an estimation procedure to be "robust" when its results are not biased by contaminants.

accuracy in parameter estimation. Some parameters (response criteria and nondecisional processes) were estimated on the basis of both target and filler trials. In this way, the authors could use several hundred trials for the parameter estimation, resulting in more stable estimates for the drift rates, which were the actual focus of their studies. Although this approach addresses the problem of sparse stimulus material, the authors were still using several hundred trials, and the question remains unanswered whether these high trial numbers are actually necessary.

There is a general consensus that higher trial numbers lead to higher accuracy in parameter estimation. This has been confirmed by several simulation studies (e.g., Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007). In these studies, however, trial numbers were manipulated only in a limited range. A more systematic comparison of different trial numbers was done by Wiecki et al. (2013; see also Ratcliff & Childers, 2015). They varied the number of trials from 20 to 150 per condition (in a design with two drift rates and s_t and s_z fixed at zero) and analyzed the mean absolute errors of the single parameters and the probability of detecting a significant difference between the two drift rates. Their results revealed an improvement in parameter estimation when the number of trials was increased.

Although these studies clearly demonstrated that parameter estimation improves with the number of trials, they did not focus on the inference of guidelines for the trial numbers required.

Method

To compare the performance of different parameter estimation methods (CS, KS, and ML) and programs (HDDM and EZ), and to infer guidelines for the numbers of trials necessary for efficient and robust parameter estimation, a set of simulation studies was carried out. In the following sections, we first describe the design of these studies, proceeding from there to present our criteria for evaluating the performance of the optimization criteria.

Design

In our studies, we tackled two different designs, in which one drift rate or two drift rates were estimated. Diffusion models with one drift rate are mostly used to analyze data that has been coded as correct (e.g., upper threshold) versus error (e.g., lower threshold). This kind of analysis allows collapsing data across different stimulus types. Alternatively, one-drift models might be applied for subsets of data based on the same stimulus types. The one-drift design was used in a first series of simulations. Both for simulation and parameter reestimation, we used models that differed in the number of free parameters. The seven-parameter model was composed of all seven parameters typically used in diffusion model analyses (a, ν, t_0, z_r, s_ν, s_{t0}, and s_{zr}); in the six-, four-, and three-parameter models, certain parameters were fixed at constant values. In the six-parameter model, the relative starting point z_r was fixed at .5 (the process starts centered between the two thresholds); in the four-parameter model, the three intertrial variabilities (s_ν, s_{t0}, and s_{zr}) were fixed at zero; and in the three-parameter model, both the intertrial variabilities and the starting point were fixed. For each model (i.e., the three-, four-, six-, and seven-parameter models), 1,000 random parameter sets were generated with typical parameter values observed in previous applications. The parameter values were drawn from uniform distributions with the minimum and maximum values shown in Table 2. Subsequently, for each parameter set, seven random data sets with different numbers of trials (24, 48, 100, 200, 500, 1,000, and 5,000) were simulated with *construct-samples*,[3] resulting in a total number of 4 (models) × 1,000 (parameter sets) × 7 (trial numbers) = 28,000 simulated data sets.

Whenever performance differs between the stimulus types, or when there is an a priori bias in favor of one of the responses, a more complex model using two drift rates is needed. Typically, thresholds are associated with responses for the two stimulus types, and drift rates are estimated separately for each stimulus type in one model (e.g., White, Ratcliff, Vasey, & McKoon, 2010a; Yap, Balota, Sibley, & Ratcliff, 2012). This results in a drift with positive sign for the stimulus at the upper threshold and a drift with negative sign for the stimulus at the lower threshold. In our simulations, this procedure was mapped by a "two-drift design." In particular, we simulated data sets with two stimulus types, using one positive and one negative drift that were allowed to vary in absolute values (i.e., difficulty). The drift values were drawn from a multivariate normal distribution. They were generated to represent a difference of d_z = 0.35 (Cohen, 1988).[4] All other parameters were equivalent to those in the one-drift design. We also used the same numbers of trials as in the one-drift design, with, for example, 24 trials composed of 12 trials of one stimulus and 12 trials of the other.[5] A total of 1,000 data sets were constructed for each of the four models (with different numbers of parameters) and for each trial number, resulting in another 28,000 data sets

[3] *Construct-samples* is part of the fast-dm software. It simulates RT data by applying a random walk with very small time steps.

[4] Effect size formula: $d_z = \frac{M_1 - M_2}{SD_{diff}}$, with $d_z = 0.35$, $M_1 = 2.00$, $M_2 = 2.35$, $SD_1 = SD_2 = 1$, and $r = .50$.

[5] CS, as implemented in fast-dm, only allows for parameter estimation if in each condition at least 12 trials are observed for one of the two responses. Accordingly, in the condition with 24 trials, the comparability of the CS method to the other estimation methods is limited, because not all datasets met this precondition.

Table 2 Minimum and maximum values of each diffusion model parameter used for the creation of parameter sets, and "possible accuracy" of each parameter

Parameter	Minimum	Maximum	Possible Accuracy[a]
a	0.5	2.0	.054
ν	–4.0	4.0	.270
t_0	0.2	0.5	.032
z_r	0.3	0.7	.035
s_ν	0.0	1.0	.849
s_{t0}	0.0	0.2	.031
s_{zr}	0.0	0.5	.402

The diffusion coefficient in fast-dm is set to 1. To compare parameter ranges and accuracies with parameter values cited in studies using a coefficient of .1, the parameters a, ν, z_r, s_ν, and s_{zr} need to be multiplied by .1. [a] 95% quantile of absolute deviations of true values and reestimated values using the ML criterion for uncontaminated simulated data sets with 5,000 trials and at least 4 % of trials at each threshold

(4 models × 7 trial numbers × 1,000 parameter sets). In the remainder of this study, we refer to the models as the *three-*, *four-*, *six-*, and *seven-parameter models*, so as to use the same terms as in the one-drift design, even if each model actually contains one further parameter (i.e., the second drift rate).

We also performed robustness tests for both the one-drift and two-drift designs. In particular, 4 % of the trials of each data set were randomly chosen and substituted for by either fast or slow contaminants. Fast contaminants were used to simulate fast guesses, that is, trials in which participants respond quickly without processing the target stimulus. Since fast-guess response is situated on the level of chance (Swensson, 1972), the response of each selected trial was randomly set to either 0 (*lower threshold*) or 1 (*upper threshold*). In terms of RTs, we used latencies at the left-hand edge of the original distribution for fast contaminants, thus ensuring that these values cannot be easily identified as outliers; real "statistical" outliers farther from the original distribution would bias the result more severely but would at the same time be easier to detect prior to analysis. More specifically, latencies for fast contaminants were drawn randomly from a uniform distribution with a range of 200 ms centered around the fastest theoretically possible time for each parameter set (i.e., using an interval from $t_{\min}$ – 100 ms to $t_{\min}$ + 100 ms, with $t_{\min} = t_0 - s_{t0}/2$). Secondly, we were interested in the influence of slow contaminants resulting from temporary distraction of participants from the task in hand. In this condition, only the RTs of the selected trials were changed, but not the respective types of response. Latencies for these types of contaminants were randomly chosen from a uniform distribution ranging from 1.5 to 5 interquartile ranges above the third quartile of the original data.

Parameter estimation

Parameter values were recovered using fast-dm-30 (Voss et al., 2015) from uncontaminated and contaminated data sets.[6] This was done using each of the three optimization criteria (CS, KS, ML). Furthermore, we analyzed all data sets with HDDM (Wiecki et al., 2013) and the data sets of the three-parameter model additionally with the EZ method (Wagenmakers et al., 2007).[7] As with our settings for HDDM (version: 0.5.3), we used 2,000 samples and a 20-sample burn-in, and the proportion of outliers was fixed at zero.[8]

In total, for both the one-drift and two-drift designs we used 28,000 data sets (4 models × 1,000 parameter sets × 7 trial numbers) with three types of contamination (none, fast, and slow) analyzed with four methods (CS, KS, ML, and HDDM), requiring 672,000 runs (+ 21,000 runs of EZ) of the parameter estimation procedure.[9]

As we mentioned before, the number of parameters reestimated was equivalent to those of the parameter model on which the simulation was based. For instance, in the case of the three-parameter model in the one-drift design, only the three parameters a, ν, and t_0 were estimated, whereas the remaining four parameters were each fixed at the correct constant value ($s_{t0} = s_{zr} = s_\nu = 0$, $z_r = .5$). The four- and seven-parameter models included estimation of starting point z_r. This is only possible if there are two distributions of responses (at the upper and lower thresholds). If no data are available for one of the two thresholds, the distance from the starting point to the "empty" threshold is not defined. Accordingly, for the models that estimated a starting point, we excluded all data sets in which the smaller distribution (response 0 or response

[6] Fast-dm was executed with the precision parameter set to 3. Setting the precision to 4 significantly slows down the estimation process without having any relevant positive impact on the parameter recovery achieved.

[7] EZ cannot be applied to data sets with an accuracy rate of 0 %, 50 %, or 100 %. For data sets with an accuracy of 100 %, we applied an edge correction method that has also been used by Wagenmakers et al. (2007): accuracy = $1-\frac{1}{2\times n}$, with n being the number of trials. We used similar approaches for the 0 % (accuracy = $\frac{1}{2\times n}$) and 50 % (accuracy = $0.5 + \frac{1}{2\times n}$) accuracy rates. In the two-drift design, EZ was applied separately to the trials of each response type. We then computed the means over the two threshold separations and the two nondecision components.

[8] The prior distributions in HDDM are a Gamma distribution (threshold separation, nondecision time), a normal distribution (drift rate, starting point), a half normal distribution (intertrial variability of drift rate and nondecision time), and a Beta distribution (intertrial variability of starting point; see also Wiecki et al., 2013). Note that these distributions differ from the one (uniform distribution for all parameters) that we used to create the parameter values, which could deteriorate the performance of HDDM.

[9] The estimations with fast-dm and HDDM were carried out using the computational resource bwUniCluster, funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

1) comprised fewer than 4 % of all trials (i.e., at least one trial at each threshold in the smallest data sets with 24 trials). The number of remaining data sets ranged from 689 to 801 out of 1,000 for the different conditions in the one-drift design. In the two-drift design, only in one condition did one data set have to be excluded due to the "4 % criterion". Because in the three- and six-parameter models the starting point is fixed (and thus the distance from a threshold without trials to the starting point is also defined), the estimation was carried out for all data sets.

Evaluation criteria

The evaluation of parameter estimation performance was based on four main criteria: (1) correlations between the true and reestimated parameter values, (2) parameter estimation biases (i.e., deviations of the reestimated from the true parameter values), (3) the numbers of participants required for detection of a drift rate difference in the two-drift design, and (4) the estimation precision, assessed as squared deviations of the reestimated from the true parameter values. In the following discussion, we present the rationale for the choice of these criteria (see also Table 3 for a summary) and give details on the computation. Additionally, (5) we evaluated the computation time required for parameter estimation.

First, parameter recovery performance was assessed by the *correlations of each parameter's true values with the reestimated values*. This criterion is of relevance if the focus of the researcher lies in the detection of relationships between diffusion model parameters and external criteria (e.g., the relationship between the drift rate and general intelligence; see, e.g., Ratcliff et al., 2010). One weakness of correlation coefficients is that they fail to reveal systematic biases in parameter recovery. Often, such a systematic bias might be unproblematic, because it does not invalidate interpretation of the results. However, there might be cases in which an estimation bias is related to the true parameter values (e.g., an estimation bias might be stronger when fewer error data are present—i.e., when drift is strong). In such a case, biased parameter estimation might challenge the internal validity of results.

Thus, our second criterion was a *measure of parameter biases*. For each parameter, we computed the differences between the estimated and true parameter values. Accordingly, a positive value indicated that the parameter was overestimated, whereas a negative value showed a parameter underestimation. Besides, we computed the mean bias for each parameter quartile (i.e., the mean of all data sets lying in the first, second, third, and fourth quartiles of the true parameter values), to graphically depict possible dependencies between the parameter values and biases. We also computed Pearson correlation coefficients between the true parameter values and the respective biases.

Note that a parameter might be estimated without bias, but still with low precision. For some participants the parameter might be overestimated, and for some underestimated, with no clear pattern. This can be a problem for difference detection, due to higher variability of the values within groups/conditions. Using a higher number of trials is one way to enhance the power of a statistical test, as parameters are estimated with less of a noise variance. Another way is to enhance the number of participants. Our third criterion was the number of participants required for the detection of a drift rate difference between two conditions. Specifically, for the two-drift model we calculated the effect sizes resulting from the recovered drift rates. Using *pwr.t.test* from the pwr R package (Champely, 2012; R Development Core Team, 2014) for the observed effect sizes between the two drift estimates, we obtained the numbers of participants required for a power of 80 % (in a two-sided paired *t* test with a significance level of 5 %). If the drift parameters were estimated perfectly (i.e., with no deviations of the estimated from the true parameter values), 66 participants would be required to detect this difference with a power of 80 % (two-sided testing).

Although an increase in the sample increases the power to detect differences between conditions, no such compensation for low precision is possible, when the aim of the researcher lies in a diagnostic application of the diffusion model (Aschenbrenner et al., 2016; Ratcliff & Childers, 2015). For this purpose, it is of great importance that parameters be estimated precisely for all individuals, thus minimizing deviations between the true and estimated values. Accordingly, our fourth evaluation criterion was the precision of parameter estimates, calculated as the *squared deviations of the reestimated from*

Table 3 Juxtaposition of the four evaluation criteria of parameter estimation performance

Evaluation Criterion	Aim of Researcher
Correlations between true and reestimated parameter values	Detection of relationships between diffusion model parameters and external criteria (e.g., between drift rate and intelligence)
Parameter estimation biases (i.e., deviations of reestimated from true parameter values)	Detection of parameter differences between conditions; interpretation of effect sizes (over- or underestimation of true effect?)
Number of participants required for detection of drift rate difference	Sample size computation for detection of parameter differences between conditions
Estimation precision—Squared deviations of reestimated from true parameter values	Diagnostic use of diffusion model parameters (e.g., drift rate for the measurement of intelligence)

the true parameter values. In contrast to the bias measure, here we did not differentiate between over- and underestimation of a parameter (using squared values, any deviation would contribute equally). Note that the diffusion model parameters have quite different scales, and the accuracy of recovery varies appreciably between parameters. Whereas, for example, t_0 can be estimated very precisely (e.g., to the third decimal place), the deviation of the true and recovered values is often much greater for the drift. Accordingly, to enhance the comparability of parameters, we standardized each parameter's bias by its respective "possible accuracy." These "possible accuracies" were deducted from an optimal parameter recovery condition—that is, from the parameter reestimations using the ML approach for data sets in the one-drift design with 5,000 trials, a minimum of 4 % of trials at each threshold, and no contaminants. From the results of these analyses, the 95 % quantiles of the absolute differences between the true and estimated parameter values were used as the "possible accuracies" for all parameters (see Table 2 for each parameter's "possible accuracy").

Finally, our last criterion—of minor importance relative to the four evaluation criteria previously presented—was *computation time*, which was the time required for the estimation process. An efficient optimization criterion should not only recover the true parameter values with high efficacy, but also require only a short time for the estimation process.

Results

In the following sections, we report our results structured by our five evaluation criteria: (1) correlations between the true and reestimated parameter values; (2) parameter estimation biases; (3) the number of participants required for detection of drift rate differences; (4) estimation precision—that is, squared deviations of the reestimated from the true parameter values; and (5) computation time.

Evaluation Criterion 1: Correlations between true and reestimated parameter values

Figure 2 shows the results obtained for our first evaluation criterion—that is, the correlations between the true and reestimated parameter values. The dependent variable in the figure is the mean correlation averaged across all parameters of the respective model using Fisher's Z transformation. The figure shows that—as expected—with higher trial numbers, higher correlation coefficients were reached. Two main aspects emanate from these correlational analyses: (1) The CS estimation criterion mostly shows lower correlation coefficients than the other estimation methods, and (2) the six- and seven-parameter models performed worse than the more restrained three- and four-parameter models. Responsible for the latter finding is the poor parameter recovery of the intertrial variability parameters s_{zr} and s_{ν}, which generally cannot be recovered well. Even under the "optimal" condition (no contamination, 5,000 trials, and ML as the optimization criterion), for s_{zr} and s_{ν} only moderate correlations of .31 and .47, respectively, are found. The performance of s_{t0} (.97) in this optimal condition is much better and, most importantly, the correlation coefficients of parameters a (1.00), t_0 (.99), ν (1.00), and z_r (.99)—which are usually of greater interest than the intertrial variabilities because of their high psychological validity (Voss et al., 2004)—are excellent.

Evaluation Criterion 2: Parameter estimation biases

Second, we analyzed parameter estimation biases. Figures 3, 4, 5, and 6 present the results of the one-drift design for the four psychologically most interesting diffusion model parameters a, ν, t_0, and z_r, respectively.[10] We will sum up the main findings from the figures, always starting with the mean bias of each parameter (indicated by the large symbols connected by lines), passing on to an examination of the relationship between the true parameter values and the biases.

As can been seen in Fig. 3, CS clearly overestimated parameter a. This overestimation decreased with the number of trials and, in the condition with no contaminants, became negligible at about 200 trials in the three- and four-parameter models, and at approximately 500 trials in the six- and seven-parameter models. The biases of the other methods were smaller and—akin to CS—became stable from around 200 to 500 trials on. In the case of slow or fast contaminants, often a notable bias in threshold separation remained even at large trial numbers. An interesting finding is observed for ML and HDDM for the condition with fast contaminants in the three- and four-parameter models. Whereas the biases of the other methods decreased with the number of trials, their biases increased (again, getting stable from around 200 to 500 trials on). This reveals that the absolute number of fast contaminants (the relative frequency was stable, with 4 % for all trial numbers) has an influence on the recovery of parameter a. We want to anticipate that a similar pattern emerged for parameter t_0, which was systematically underestimated by ML and HDDM, with this bias increasing with the number of trials. This makes sense, because these methods try to account for all RTs and adapt t_0 to the smallest observed time. With the inclusion of s_{t0}—as in the six- and seven-parameter models—the biases were much smaller, because s_{t0} helps to explain very fast RTs.

Next, we analyzed whether and how the bias depends on the true value of the parameter. For the condition with no contaminants, there were at maximum small relationships with no clear pattern ($|r| < .30$). For the condition with slow contaminants, however, the relationship of the true parameter

[10] We also analyzed biases for the two-drift design. The findings were very similar to those from the one-drift design.

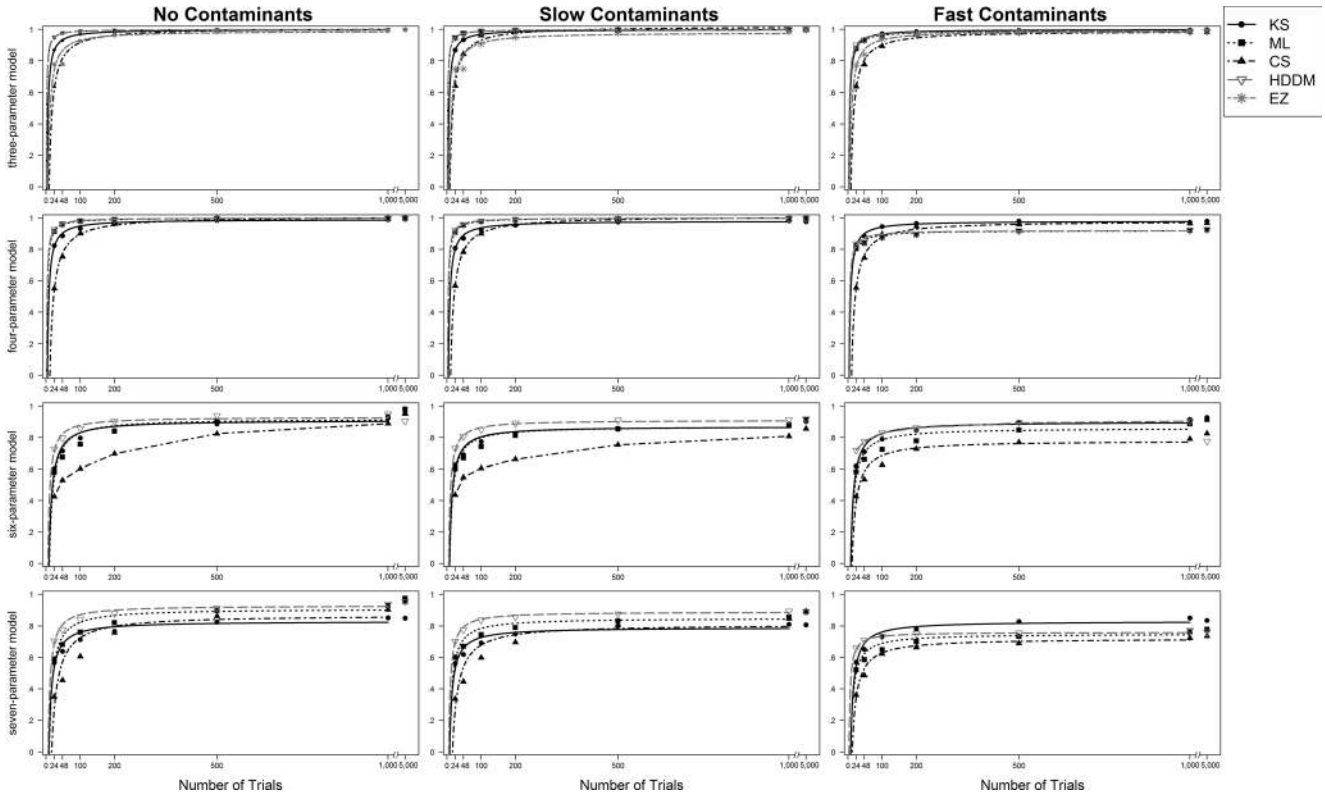

Fig. 2 Scatterplot of mean correlation between true and reestimated parameters in the one-drift design, for uncontaminated data sets (left column), data sets with slow contaminants (middle column) and data sets with fast contaminants (right column). On the basis of data sets with at least 4 % of trials at each threshold. Power functions were fitted to the data. Whenever the curve was a poor fit ($R^2 < .80$), lines were drawn between adjacent trial numbers

value *a* to the bias increased with the number of trials. For example, for the three-parameter model estimated by ML, the correlation rose from $r = -.07$ to .89, for $n = 24$ and $n = 5{,}000$, respectively. This increase with the number of trials was less pronounced for the more complex models (e.g., for the seven-parameter model and ML: $r = .15$ at $n = 24$ and $r = .46$ at $n = 5{,}000$). In the condition with fast contaminants, the pattern was less clear-cut. For KS and CS, there were mostly no relationships or very small relationships. For ML and HDDM, on the other hand, especially in the three- and four-parameter models, a (negative) relation of the true value and the bias increased with the trial number (e.g., for the four-parameter model and HDDM: $r = -.08$ at $n = 24$ and $r = -.64$ at $n = 5{,}000$).

Akin to parameter *a*, for the drift rate, biases (mostly overestimation, especially in the six- and seven-parameter models) got stable at approximately 200–500 trials. Relationships between the true drift value (with negative true values transformed into positive values[11]) and the respective bias were very small for data with no contaminants in basically all conditions. For data with slow contaminants, the (negative) relationship increased with the number of trials, especially in the three- and four-parameter models (e.g., for the four-parameter model and ML: $r = -.30$ for $n = 24$ and $r = -.95$ for $n = 5{,}000$). A similar increase was observed for the condition with fast contaminants in the three-parameter model, and for ML and HDDM in the six-parameter model. In the four- and seven-parameter models, the relationships were mostly positive, with a smaller influence of the number of trials.

The nondecision time was estimated quite precisely in the conditions with no or slow contaminants. Again, stability of the biases was reached at 200–500 trials. In the condition with fast contaminants, we observed a systematic underestimation, which is plausible given the added fast outliers. As we mentioned before, for ML and HDDM in the three- and four-parameter models, this bias increased essentially with the number of trials. Importantly, there was no relevant relationship between the true value of t_0 and the sign and size of the bias (with the exception of HDDM showing negative correlations of at maximum $r = -.28$; these relationships decreased with the number of trials).

Finally, the pattern for z_r revealed that this parameter was more often under- than overestimated. More importantly, we

[11] This transformation was used so that the four quartiles would span a range from a very slow to a very high speed of information accumulation. Note that the results were very similar if positive and negative true drift rates were analyzed separately.

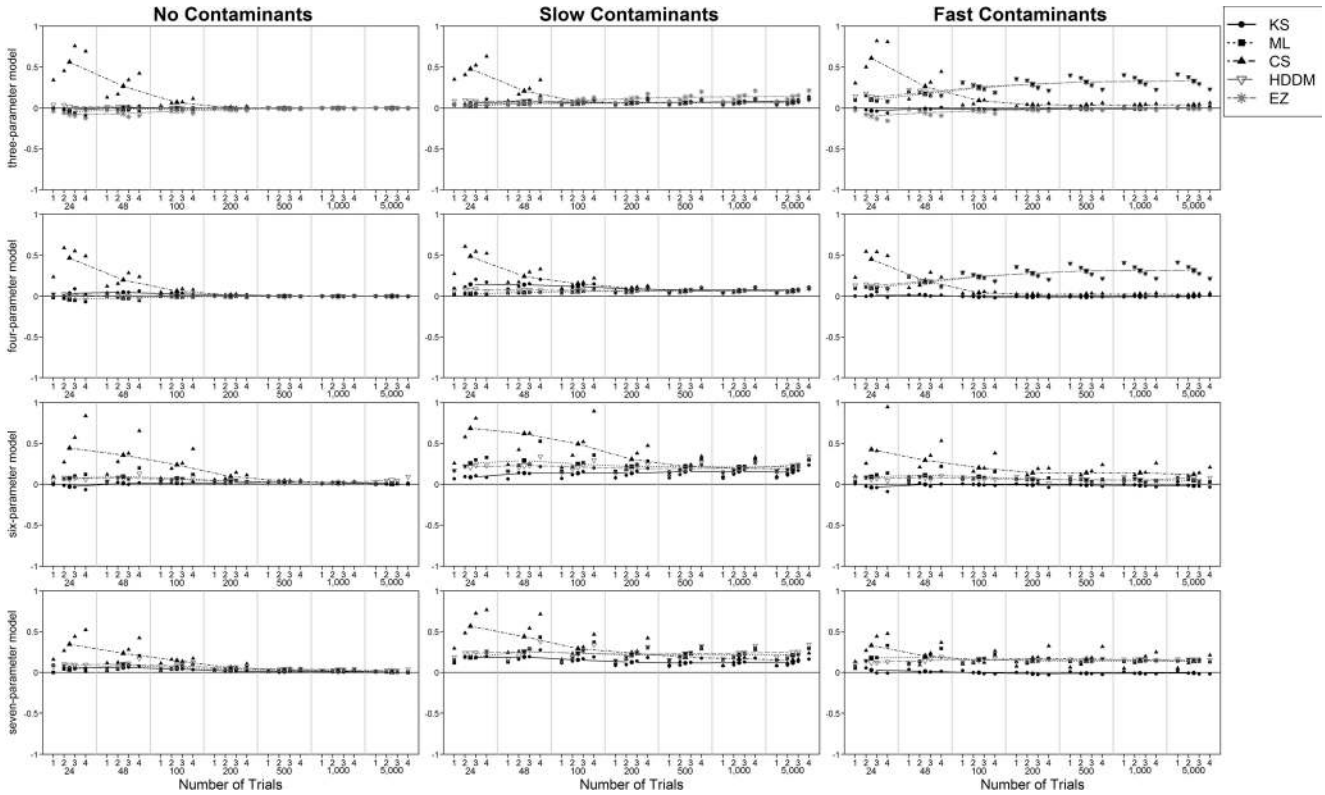

Fig. 3 Mean differences between estimated and true values of *parameter a* for each quartile of the true parameter values (numbers 1–4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. On the basis of data sets with at least 4 % of trials at each threshold. Few values are not depicted as they fall outside the *y*-axis limits

found a negative relationship between the size of z_r and the bias present for almost all conditions. There was also no clear improvement with the number of trials. Sometimes the relationship decreased in absolute value (e.g., for no contaminants in the seven-parameter model and HDDM: $r = -.70$ for $n = 24$ and $r = -.10$ for $n = 5{,}000$); often, however, it did not change, or even increased (e.g., for no contaminants in the seven-parameter model and KS: $r = -.29$ for $n = 24$ and $r = -.38$ for $n = 5{,}000$). The method with the smallest absolute correlation was ML. However, in the condition with fast contaminants, all methods featured essential negative relationships (e.g., for the seven-parameter model, ML: $r = -.45$ for $n = 5{,}000$).

Evaluation Criterion 3: Numbers of participants required for detection of a drift rate difference

Figure 7 shows the numbers of participants required for detecting a difference in drift rates ($d_z = 0.35$) in a two-sided paired *t* test with a power of .80 conditional on the number of trials. If parameters were recovered perfectly (i.e., the estimated drift rates were equivalent to the true drift rates), 66 participants were needed for the detection of this difference (represented by the horizontal line in the figure). Obviously, parameters are estimated less precisely from small than from higher trial numbers. Thus, more participants are required in order to compensate for the inflated error variance. Figure 7 shows that an increase of trial numbers above 200–500 did not further reduce the required sample size. In most conditions, ML outperformed the other methods. Interestingly, even for data sets with fast contaminants, ML showed a good performance. Furthermore, HDDM failed to outperform the non-Bayesian ML approach in either condition. A further finding is that the performance of EZ was generally very good.

Evaluation Criterion 4: Estimation precision—Squared deviations of reestimated from true parameter values

Akin to the mean correlation coefficient over all parameters, we also computed an average measure for the squared deviations. Figure 8 shows the 95 % quantiles of these mean squared deviations for each condition, depending on the number of trials in the one-drift design. The use of the 95 % quantiles makes it possible to compare the worst cases for each condition, since deviations are smaller for most data sets. If the parameters are to be used for diagnostic purposes, it is important that the parameters be estimated accurately for all individuals.

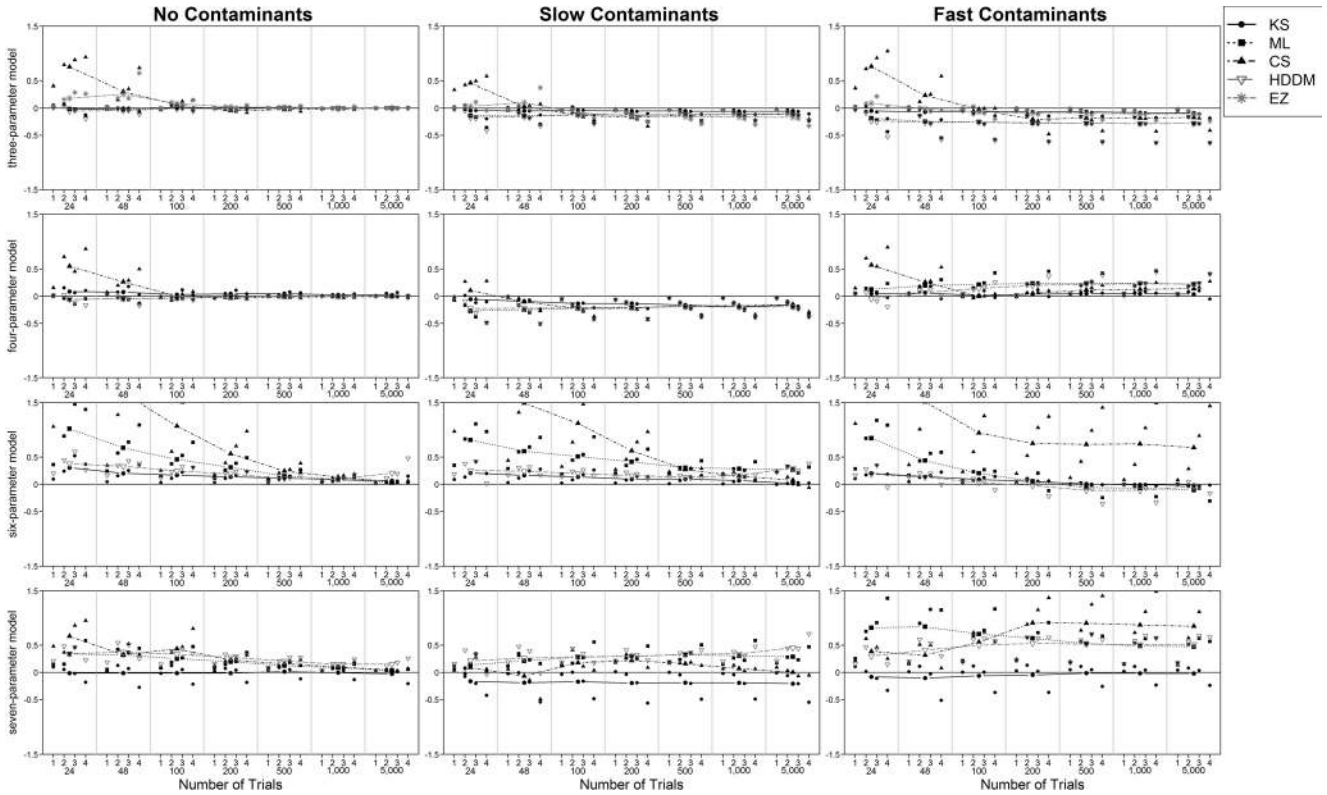

Fig. 4 Mean differences between estimated and true values of *parameter* ν for each quartile of the true parameter values (numbers 1–4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. All negative drift values were transformed to positive values so that the true values are all located between 0 and 4. On the basis of data sets with at least 4 % of trials at each threshold

One central aim of this article is to provide guidelines on the numbers of trials required for diffusion model analyses. Because the squared deviation criterion is the strictest criterion, we used this criterion for the definition of required trial numbers.

In the subsequent section, we first specify our procedure for identifying the trial numbers required. Then we report the results for data sets without contaminants, followed by the results observed in the conditions with contaminants. Whereas Figure 8 shows the "mean deviations" (averaged over all parameters of the respective model), in the following sections we present results separately for the four main diffusion model parameters (i.e., a, ν, t_0, and z_r).

Criteria for trial numbers required As can be seen from Fig. 8, for the one-drift design, the higher the number of trials, the better the estimation usually is.[12] For uncontaminated data, the relation of deviations to trial numbers is mostly described well by power functions.[13] To find the requisite trial numbers, the fitted power functions were used whenever they fitted well (i.e., when the adjusted R^2 was at minimum .80); otherwise, linear interpolation was used.

We defined a squared deviation of 15 as a criterion for the minimal number of trials required, indicating that the 95 % quantiles of the deviations should be no more than 15 times as large as in the "optimal" condition. This value is obviously quite high (allowing for large deviations), and at least in part arbitrary. For interpretation, one has to bear in mind that 95 % of data sets would fit better (i.e., have a squared deviation below 15). We further determined the number of trials at which the stricter criterion of deviations of 5 was reached for 95 % of the data sets, thereby deriving guidelines on the trial numbers needed for low (15) and high (5) precision.

As the asymptotic courses of the fitted functions describing the relation of trial numbers to mean deviations in Fig. 8 illustrate, adding further trials is very helpful when the number of trials is small, but for higher trial numbers further increases bring only marginal gains in accuracy. Accordingly, we also defined the number of trials above which a further increase

[12] Some exceptions have been found. We observed that the performance of KS deteriorated from the condition with 1,000 to that with 5,000 trials in the four- and seven-parameter models. So did the performance of HDDM in the six- and seven-parameter models. We had similar findings using the two-drift design.

[13] $D = b_0 \cdot n^{b_1}$, where D is the 95 % quantile of the squared deviations and n is the number of trials.

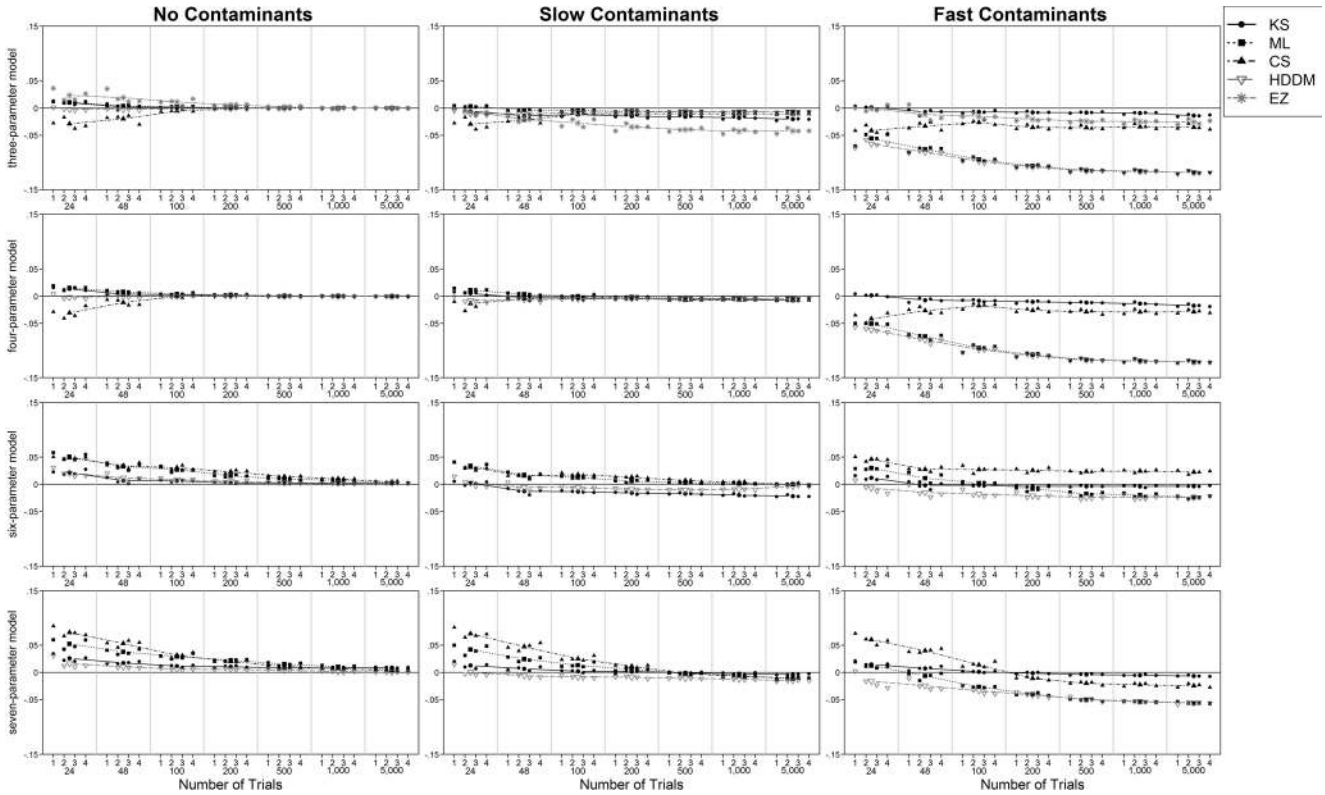

Fig. 5 Mean differences between estimated and true values of *parameter* t_0 for each quartile of the true parameter values (numbers 1–4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. On the basis of data sets with at least 4 % of trials at each threshold. Few values are not depicted as they fall outside the *y*-axis limits

had only a minimal impact on the quality of parameter recovery. As a criterion, we used the point at which the functions describing the relation of deviations to trial numbers had a slope of –0.01. The trial numbers required for low and for high precision and the limit at which a further increase made little sense are presented in Table 4 (one-drift model; at least 4 % of trials at each threshold), Table 5 (one-drift model; less than 4 % of trials at one threshold), and Table 6 (two-drift model; at least 4 % of trials at each threshold). Trial numbers are given separately for the four main diffusion model parameters (a, ν, t_0, and z_r), depending on the complexity of the parameter model (three-/four-/six-/seven-parameter models), the type of contamination (none/fast/slow), and the estimation method (KS/ML/CS/HDDM/EZ).

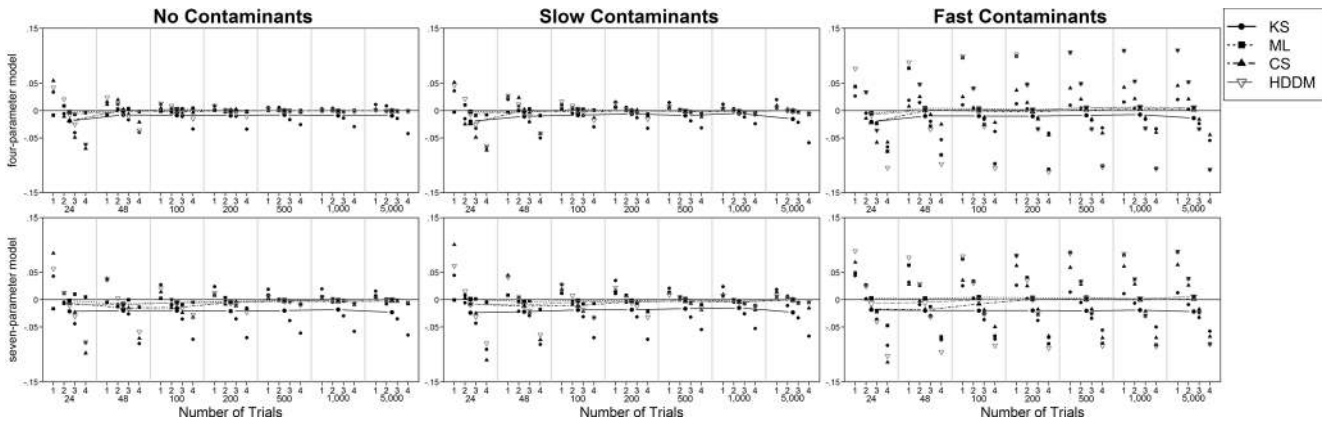

Fig. 6 Mean differences between estimated and true values of *parameter* z_r for each quartile of the true parameter values (numbers 1–4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. On the basis of data sets with at least 4 % of trials at each threshold. Few values are not depicted as they fall outside the *y*-axis limits

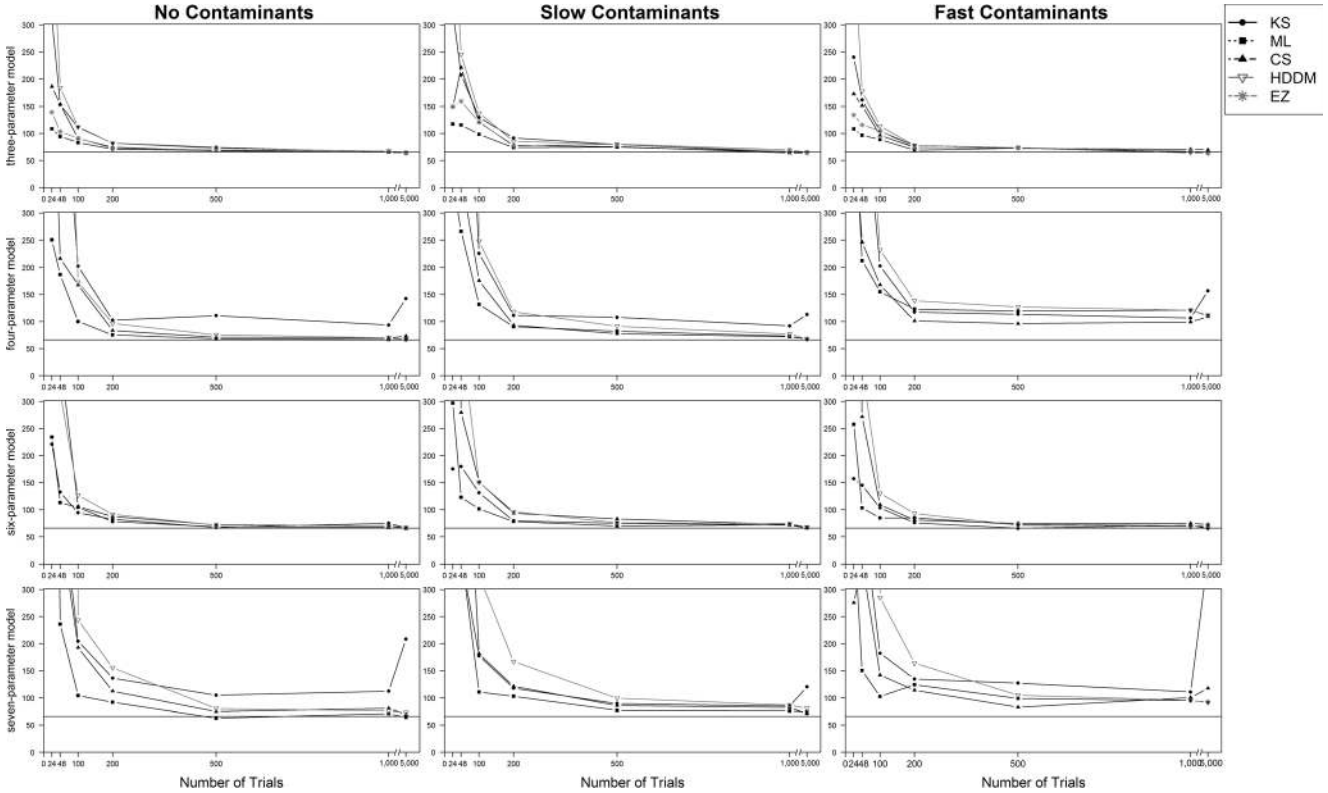

Fig. 7 Scatterplot of the number of participants required for the detection of a difference in reestimated drift rates, depending on the number of trials and the estimation method. The horizontal line indicates the number of participants required for the original effect size ($n = 66$ for $d_z = 0.35$). On the basis of data sets with at least 4 % of trials at each threshold. Required numbers of participants exceeding 300 are not depicted

Trial numbers required for uncontaminated data In the three- and four-parameter models of the one-drift design, using ML or HDDM, low precision could be reached with fewer than 60 trials, and high precision with fewer than 160 trials. KS also performed well, with fewer than 200 trials for low precision. EZ applied to the three-parameter model was competitive with ML, HDDM, and KS in terms of drift rate estimation, with approximately 70 trials for low and 200 trials for high precision. Parameters a and t_0, on the other hand, were estimated worse. CS showed the poorest performance requiring still fewer than 290 trials for low precision. The comparison of the different parameters reveals that with the exception of EZ, the nondecision time required the least number of trials, followed by the drift rate, and the threshold separation. Parameter z_r was estimated very well by ML and HDDM (requiring fewer than 40 trials for low precision), whereas KS and CS required more trials (< 170).

In the six- and seven-parameter models, more trials were required than in the three- and four-parameter models. Again, the lowest trial numbers were always needed for the nondecision time, and the highest numbers were usually required for the threshold separation. The drift rate was estimated best by KS, with fewer than 200 trials for low precision in both models. CS, on the other hand, required more than 700 trials in the six-, and more than 400 trials in the seven-parameter model. In fact, CS is usually applied with such trial numbers (or even higher ones), and should thus give reliable results. However, our results also show that other methods can supply satisfying reliability already with smaller trial numbers.

For data sets with fewer than 4 % of trials at one of the two thresholds, the estimation of parameters a and ν requires higher trial numbers (see Table 5). Only t_0 was estimated with a performance similar to that for the other data sets.

The numbers of trials required by the two-drift design are depicted in Table 6, analogous to Table 4 for the one-drift design.[14] The parameter that suffered most from the more complex design was the threshold separation. The drift rate was also estimated worse, whereas there was no deterioration (and sometimes even an improvement) for the nondecision time. Besides, the starting point was estimated better in the two-drift design. As for the comparison of the different optimization criteria, the pattern was similar to the one observed for the one-drift design. Most importantly, HDDM in sum showed the best performance, followed by ML, KS, and CS. EZ also performed very well, even beating ML for the estimation of the drift rate.

[14] The requisite trial numbers for the drift rate are based on the mean squared deviations of the two drift rates.

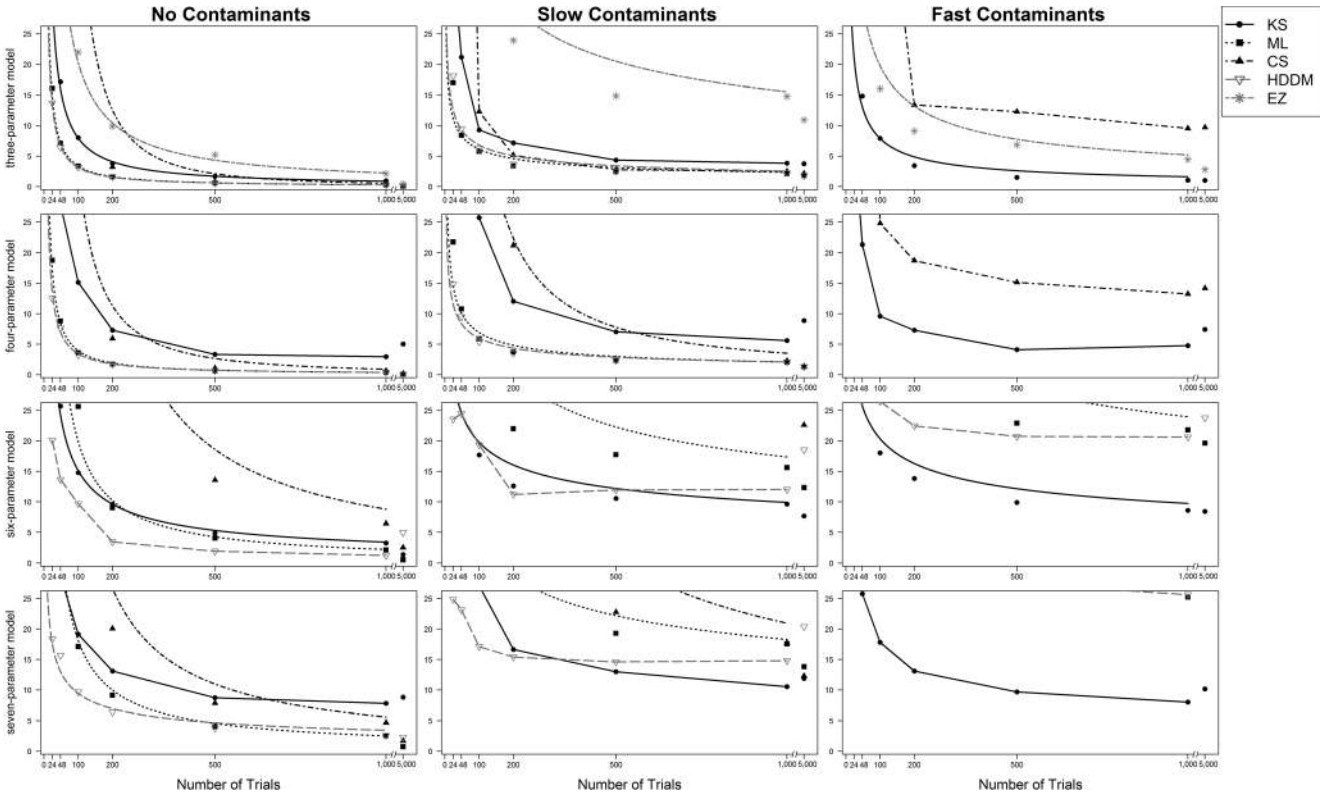

Fig. 8 Scatterplot of 95% quantiles of mean deviation between true and reestimated parameters in the one-drift design, for uncontaminated data sets (left column), data sets with slow contaminants (middle column) and data sets with fast contaminants (right column). On the basis of data sets with at least 4 % of trials at each threshold. Quantiles exceeding a mean deviation of 25 are not depicted. Power functions were fitted to the data. Whenever the curve was a poor fit ($R^2 < .80$), lines were drawn between adjacent trial numbers

For a better understanding of the precision of the results for trial numbers derived from the criteria for low and high precision, we also calculated the *correlations of the true and recovered parameters* at these points. Toward this aim, power functions[15] or linear interpolation (if the adjusted R^2 was beneath .80) were used. Importantly, the correlations were generally very high (most of them above .90), and therefore do not imply that even stricter criteria should be applied for the trial numbers required. The correlation coefficients were lowest for parameters t_0 (ranging from .79 to .97) and z_r (.76–.96). Note that the trial numbers required for t_0 were often very low (even $n < 24$). Because usually many more trials will be used, even higher correlation coefficients may be reached.

Besides the requisite trial numbers, Tables 4, 5, and 6 also show the maximum trial numbers based on the slope criterion. For instance, in the three- and four-parameter models with uncontaminated data and at least 4 % of trials at each threshold, HDDM and ML reached this criterion for all parameters after fewer than 300 trials in the one-drift design, and after fewer than 600 trials in the two-drift design. In the six- and seven-parameter models, the criterion was reached after fewer than either 700 trials (one-drift design) or 1,000 trials (two-drift design). Generally, the criterion was reached earlier for nondecision time, drift rate, and starting point than for threshold separation.

Trial numbers required for contaminated data Up to this point, we have only presented results for the condition without contaminant trials. The middle and right columns of Fig. 8 show the mean deviations of the recovered parameters from contaminated data. As can be seen in Tables 4 (one-drift design) and 6 (two-drift design), in the condition with *slow contaminants*, parameter *a* was estimated much worse, requiring more trials in almost all conditions. The drift rate did not suffer much in the three- and four-parameter models, but it often required many more trials in the six- and seven-parameter models. The nondecision time and starting point often did not suffer from the addition of slow contaminants. Finally, EZ estimated the drift rate quite well, but nondecision time and, especially, threshold separation required much higher trial numbers than in the condition with no contaminants and than the other methods.

In the presence of *fast contaminants*, KS continued to display good parameter recovery. By contrast, the results from both ML and HDDM were affected strongly by the occurrence of fast contaminants. This applied to both the threshold

[15] That is, correlation = $b_0 + b_1/n$, where n is the number of trials.

Table 4 Numbers of trials required in the one-drift design for data sets with at least 4 % of trials at each threshold, depending on the parameter model, estimated parameter, type of contamination, and estimation method

		Three-Parameter Model			Four-Parameter Model				Six-Parameter Model			Seven-Parameter Model			
		a	ν	t_0	a	ν	t_0	z_r	a	ν	t_0	a	ν	t_0	z_r
No contaminants	KS	125; 403	37; 119	26; 83	199; 589	94; >5,000	<24; 78	163; >5,000	223; 1,122	102; 479	<24; 91	650; 3,616	195; >5,000	<24; 310	>5,000
		436	241	203	546	500	195	500	551	402	199	824	500	217	500
	ML	55; 158	26; 78	<24; <24	57; 158	42; 126	<24; <24	38; 109	272; 783	277; 727	56; 223	318; 1,012	300; 944	54; 335	132; 548
		283	197	89	286	251	99	235	641	649	309	687	668	313	449
	CS	287; 498	105; 219	50; 104	259; 478	149; 352	56; 117	146; 348	759; 1,693	713; 1,609	94; 421	592; 1,397	453; 1,276	124; 517	278; 816
		628	358	229	599	457	246	452	1,156	1,112	388	992	833	438	647
	HDDM	48; 142	26; 79	<24; <24	53; 152	35; 112	<24; <24	26; 85	>5,000	94; >5,000	<24; 35	378; 2,788	288; 4,488	<24; 51	41; 290
		267	198	86	277	235	86	206	500	500	134	626	486	150	284
	EZ	272; 836	68; 201	109; 354											
		638	318	408											
Slow contaminants	KS	555; >5,000	28; 128	38; 147	1,031; >5,000	81; >5,000	<24; 106	>5,000	>5,000	94; 605	28; 253	>5,000	>5,000	39; 195	>5,000
		644	234	200	997	200	188	500	968	385	100	1,411	200	200	1,000
	ML	116; 1,945	<24; 116	<24; <24	135; 1,963	39; 508	<24; <24	41; 224	>5,000	1,493; >5,000	30; 257	>5,000	>5,000	<24; 350	218; 2,801
		372	201	74	395	278	75	281	1,987	1,047	256	1,925	1,019	215	465
	CS	192; >5,000	83; 292	47; 91	719; 1,870	186; 722	57; 96	211; 564	>5,000	1,329; 4,032	84; 1,310	>5,000	791; 3,355	146; 1,074	375; 1,150
		500	362	200	1,081	525	200	562	4,593	1,405	342	3,202	973	446	748
	HDDM	143; 2,230	<24; 121	<24; <24	146; 1,881	24; 427	<24; <24	<24; 133	>5,000	68; >5,000	<24; 30	>5,000	>5,000	<24; 40	29; 1,475
		398	204	86	413	238	78	211	1,000	200	100	500	200	100	234
	EZ	>5,000	69; 475	1,246; >5,000											
		1,208	343	587											
Fast contaminants	KS	109; 402	26; 133	29; 92	144; 540	89; >5,000	<24; 99	>5,000	232; 1,407	69; 302	<24; 110	199; >5,000	245; >5,000	29; 149	>5,000
		412	231	200	467	500	200	500	544	340	205	500	500	100	500
	ML	>5,000	45; >5,000	>5,000	>5,000	>5,000	>5,000	>5,000	578; 2,659	287; 1,155	28; 341	>5,000	>5,000	<24; >5,000	>5,000
		48	100	100	100	500	200	100	831	634	252	1,841	1,814	100	200
	CS	418; >5,000	146; >5,000	694; >5,000	217; >5,000	100; >5,000	87; >5,000	>5,000	>5,000	>5,000	132; 3,040	>5,000	>5,000	97; >5,000	>5,000
		500	200	100	500	200	100	500	1,000	500	367	1,000	1,000	200	100
	HDDM	>5,000	47; >5,000	>5,000	>5,000	>5,000	>5,000	>5,000	177; >5,000	57; >5,000	<24; 38	>5,000	>5,000	<24; >5,000	>5,000
		24	100	100	100	500	200	24	500	200	100	500	200	100	24
	EZ	257; 923	46; 269	296; >5,000											
		612	294	409											

The cells comprise the requisite trial numbers for low and high precision (first row) and the limit (i.e., the number of trials not worth exceeding, since performance then improves only marginally; second row)

Table 5 Numbers of trials required in the one-drift design for data sets with fewer than 4 % of trials at one threshold, depending on the parameter model, estimated parameter, type of contamination, and estimation method

		Three-Parameter Model			Six-Parameter Model		
		a	ν	t_0	*a*	ν	t_0
No contaminants	KS	>5,000 2,086	101; 473 400	35; 147 253	3,813; >5,000 1,316	213; 1,918 493	<24; 94 202
	ML	463; 1,003 870	74; 183 314	<24; 47 148	1,584; 3,491 1,777	760; 1,516 1,182	41; 126 250
	CS	1,703; 3,492 1,916	206; 502 550	102; 221 358	4,988; >5,000 >5,000	4,949; >5,000 >5,000	36; 163 259
	HDDM	439; 1,275 816	48; 142 267	<24; 25 112	>5,000 1,000	>5,000 500	<24; 32 130
	EZ	>5,000	4,087; >5,000 500	<24; 4,544 1,000			
Slow contaminants	KS	>5,000 500	98; >5,000 200	<24; 102 192	>5,000 1,000	316; >5,000 492	33; 76 100
	ML	603; 2,965 830	44; 420 100	<24; <24 91	4,821; >5,000 2,397	505; 1,114 914	<24; 142 214
	CS	>5,000 500	298; >5,000 500	87; 100 200	>5,000	4,938; 4,997 >5,000	<24; 175 229
	HDDM	568; 2,928 799	32; 356 100	<24; <24 91	>5,000 1,000	>5,000 500	<24; <24 77
	EZ	>5,000 1,000	4,295; >5,000 500	<24; >5,000 500			
Fast contaminants	KS	>5,000 500	124; >5,000 200	<24; 175 200	>5,000 500	180; 1,547 469	<24; 113 182
	ML	>5,000 1,000	100; 3,465 200	>5,000 200	>5,000	>5,000	71; 99 200
	CS	3,417; >5,000 2,653	>5,000 200	>5,000 200	>5,000 1,000	>5,000 1,000	46; 394 100
	HDDM	>5,000 4,762	80; 1,840 322	>5,000 283	>5,000 1,000	>5,000 200	<24; 28 98
	EZ	>5,000 1,000	498; >5,000 1,000	<24; >5,000 500			

The cells comprise the requisite trial numbers for low and high precision (first row) and the limit (i.e., the number of trials not worth exceeding, since performance then improves only marginally; second row).

separation and nondecision component in the three- and four-parameter models, and to all parameters in the four-parameter model. Here, ML and HDDM stayed above the critical value of 15. In the more complex models, nondecision time was estimated better (probably due to the intertrial variability of nondecision time; see also the findings for the bias measure). Besides, in the six-parameter model, especially, HDDM turned in a good performance for the other parameters as well. Across the different parameter models, the performance of CS was often better than that of ML and HDDM, but still worse than that of KS. Interestingly, EZ showed a good performance for drift rate and nondecision time despite the presence of fast contaminants. For the data sets in which the smaller response distribution comprised fewer than 4 % of the data, the pattern of results was similar.

Since the added contaminant trials were situated partly outside, partly overlapping with the RT distribution, they could not all be identified and excluded before parameter estimation. Applying the frequently used criterion of 200 ms as the lower limit for the condition with fast contaminants to the one-drift design led to the exclusion of 0.6 % of the trials on average (so only a small part of the 4 % contaminants were identified). We additionally applied the Tukey criterion (Tukey, 1977) to exclude further possible contaminants. In the condition with fast contaminants, this led to a total exclusion of 5.5 % of the trials on average. The average percentage of trials correctly identified as fast contaminants was 98.4 %. However, also 4.7 % of the trials were falsely identified as slow contaminants. In the condition with slow contaminants, 7.1 % of the slow trials were excluded, but only 56.3 % of these were "true" slow contaminants (the percentage of falsely identified fast contaminants was very small).

To see whether the exclusion of trials led to an improvement in parameter recovery, we reestimated the parameters for

Table 6 Numbers of trials required in the two-drift design for data sets with at least 4 % of trials at each threshold, depending on the parameter model, estimated parameter, type of contamination, and estimation method

		Three-Parameter Model			Four-Parameter Model				Six-Parameter Model			Seven-Parameter Model			
		a	ν	t_0	a	ν	t_0	z_r	a	ν	t_0	a	ν	t_0	z_r
No contaminants	KS	1,056; 3,680	121; 389	39; 122	1,010; >5,000	152; >5,000	<24; 52	95; >5,000	1,234; >5,000	341; 1,240	41; 156	863; >5,000	295; >5,000	<24; 103	175; >5,000
		1,189	429	245	791	500	157	200	1,122	695	266	848	500	210	500
	ML	201; 509	59; 161	<24; <24	171; 434	83; 226	<24; <24	29; 85	536; 1,348	498; 1,091	43; 135	412; 1,149	454; 1,043	32; 101	64; 175
		545	289	100	500	345	93	207	928	907	258	794	856	223	301
	CS	816; 1,535	183; 398	72; 150	677; 1,255	226; 545	64; 129	53; 180	1,790; 3,443	1,504; 3,260	82; 280	894; 1,833	761; 1,582	45; 146	103; 289
		1,248	505	286	1,109	579	260	292	2,034	1,736	358	1,298	1,174	266	388
	HDDM	155; 454	40; 124	<24; <24	143; 416	59; 182	<24; <24	<24; 64	369; >5,000	127; 479	<24; 44	>5,000	>5,000	<24; 34	28; 111
		481	248	86	462	299	82	179	500	500	150	500	500	134	227
	EZ	295; 1,012	49; 160	28; 93											
		655	278	214											
Slow contaminants	KS	1,974; >5,000	122; 736	26; 178	>5,000	133; >5,000	<24; 48	96; >5,000	>5,000	350; 1,646	32; 331	>5,000	402; 3,964	28; 84	276; 2,854
		1,318	426	240	200	200	200	500	1,872	671	262	500	596	100	522
	ML	261; 1,452	47; 279	<24; <24	227; 1,263	90; >5,000	<24; <24	26; 120	>5,000	1,533; >5,000	<24; 139	>5,000	1,277; >5,000	<24; 91	50; 216
		577	200	80	547	200	70	226	2,214	1,306	223	1,896	1,138	195	296
	CS	1,039; 2,596	245; 907	42; 137	798; 1,847	288; 1,065	50; 91	52; 188	>5,000	2,143; >5,000	79; 396	4,358; >5,000	1,083; 2,473	31; 155	101; 333
		1,334	597	258	1,179	642	200	292	4,357	2,004	362	2,815	1,407	247	394
	HDDM	255; 1,692	34; 570	<24; <24	231; 1,320	82; 2,301	<24; <24	<24; 71	>5,000	>5,000	<24; <24	>5,000	>5,000	<24; <24	<24; 135
		554	264	85	548	317	78	174	1,000	200	88	1,000	200	88	216
	EZ	>5,000	93; 2,595	117; >5,000											
		1,000	327	195											
Fast contaminants	KS	1,876; >5,000	126; 667	30; 132	1,142; >5,000	153; >5,000	<24; 49	96; >5,000	1,991; >5,000	275; 970	38; 163	1,393; >5,000	388; 3,496	<24; 75	170; >5,000
		1,291	437	239	780	500	152	500	1,110	632	264	829	602	179	500
	ML	>5,000	86; >5,000	>5,000	>5,000	>5,000	>5,000	>5,000	2,254; >5,000	1,148; 3,301	<24; 264	1,309; 4,322	1,329; >5,000	33; 94	378; >5,000
		500	200	200	1,000	200	200	200	1,839	1,334	229	1,346	1,256	200	200
	CS	2,425; >5,000	>5,000	>5,000	2,045; 4,857	>5,000	>5,000	>5,000	>5,000	>5,000	87; 1,574	>5,000	>5,000	40; 136	>5,000
		2,163	200	200	1,982	200	100	100	500	4,906	340	200	500	200	200
	HDDM	>5,000	90; >5,000	>5,000	>5,000	>5,000	>5,000	>5,000	>5,000	78; >5,000	<24; 44	294; >5,000	193; >5,000	<24; 57	408; >5,000
		1,122	200	131	843	200	200	200	500	200	100	500	200	100	329
	EZ	>5,000	48; 473	<24; 102											
		500	200	194											

The cells comprise the requisite trial numbers for low and high precision (first row) and the limit (i.e., the number of trials not worth exceeding, since performance then improves only marginally; second row)

the adjusted data sets. This procedure led to basically the same results as when all trials were used for parameter estimation. For almost all cases in the condition with fast contaminants, the numbers of trials required were equal to or higher than the values with the full data set. For data sets with slow contaminants, an improvement was observed for some conditions (mostly in the six- and seven-parameter models), but a deterioration or equal performance in most conditions. In sum, no systematic overall improvement pattern could be identified from the exclusion of trials according to the standard procedure of identifying outliers.

Another option for dealing with possible contaminant trials would be including in the model a further parameter to explicitly estimate the percentage of contaminant trials. To exemplify the effect of this additional parameter, we implemented the approach proposed by Ratcliff and Tuerlinckx (2002) to the ML criterion. The requisite trial numbers resulting from the inclusion of this further parameter were compared to the trial numbers shown in Tables 4 and 5. For the data sets with slow contaminants, we observed improvements for some conditions (almost all of them in the six- and seven-parameter models), but deteriorations in other conditions (in the three- and four-parameter models). For the conditions with fast contaminants, the criteria for low and high precision were—as for the data without adjustments—mostly not reached.[16] In total, the inclusion of this further parameter, at least for the range of trial numbers analyzed in our study, did not have a clear positive effect. The positive effect possibly resulting from the estimation of the proportion of contaminants might have been undermined by the negative effect of adding a further parameter calling for estimation.

Criterion 5: Computation time

Our final evaluation criterion was the computation time needed for parameter estimation, averaged across individual data sets. In the three- and four-parameter models, no relevant time difference was apparent between the three optimization criteria KS, ML, and CS. On average, parameter estimation took less than 5 s per individual data set for all methods. Only after inclusion of the intertrial variabilities did the three optimization criteria differ substantially in terms of computation time, with ML for large trial numbers taking considerably longer than KS and CS. Even then, however, the computation process took no longer than 30 min per data set in the one-drift design, and 40 min in the two-drift design. Accordingly, as long as the traditional methods are used, computation time will probably not affect a researcher's choice of optimization criterion. The HDDM approach, however, involved longer computation times, requiring anything up to 5 h per data set. Because EZ is based on closed-form equations, the computation time was negligibly small.

[16] HDDM also permits estimation of the proportion of contaminants, using an approach similar to the one used by Ratcliff and Tuerlinckx (2002). We applied this approach to our data and found results very similar to those observed for the non-Bayesian ML approach. Most importantly, the inclusion of the additional parameter did not have a consistent positive effect on parameter estimation.

Discussion

As is demonstrated by the increasing number of research articles applying Ratcliff's diffusion model (Ratcliff, 1978), the interest in diffusion modeling is growing in various fields of psychology (Voss et al., 2013). This development can be attributed to a recognition of the main benefit of the diffusion model, that is, its capacity to disentangle several latent cognitive processes. The recent increase in popularity of the diffusion model is further fostered by the availability of user-friendly software solutions. Due to these programs the growing interest in diffusion modeling is not hampered by any lack of mathematical or programming skills (Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2007; Wagenmakers et al., 2007). However, knowledge is still scarce about the preconditions of diffusion modeling. In any diffusion model study, the probably most important issue that a researcher has to examine is validity of parameters. In recent years, several experimental validation studies (e.g., Arnold, Bröder, & Bayen, 2015; Voss et al., 2004; Wagenmakers, Ratcliff, et al., 2008) and correlational analyses (e.g., Ratcliff, Thapar, & McKoon, 2011; Schubert, Hagemann, Voss, Schankin, & Bergmann, 2015) have supplied promising results regarding parameter validity. However, for any new paradigm, the validity has to be first examined.

A second prerequisite for diffusion modeling is robustness of parameter estimation. One important question here regards the amount of data that are required. Typically, very large numbers of trials (>1,000) have been used in diffusion model analyses (e.g., Ratcliff & Rouder, 1998; Wagenmakers, Ratcliff, et al., 2008). The present article aimed at clarifying whether this convention could be corroborated by data. Only very few studies have systematically analyzed the effects of different numbers of trials on the precision of parameter estimation (e.g., Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009). To fill this gap, we ran a set of simulation studies using different numbers of trials with the aim of deducing guidelines for the necessary trial numbers. In these studies, the precision of parameter estimation was compared for models differing with regard to the number of parameters while using different optimization criteria. In particular, we analyzed parameter recovery for three-parameter (a, ν, t_0), four-parameter (a, ν, t_0, z_r), six-parameter (a, ν, t_0, s_ν, s_{t0}, s_{zr}), and seven-parameter (a, ν, t_0, z_r, s_ν, s_{t0}, s_{zr}) models, with either one drift rate or two different drift rates. Data sets were simulated either without contaminated trials or with 4 % of slow or fast contaminants. Then, the parameters were reestimated using the KS, ML, and CS methods, as well as a Bayesian approach (HDDM; Wiecki et al., 2013). Besides, the

EZ-diffusion model (Wagenmakers et al., 2007) was applied to the data of the three-parameter model.

Parameter estimation performance was evaluated using different criteria. (1) First, we analyzed correlations between true and reestimated parameters, which is of relevance for researchers interested in relationships between diffusion model parameters and external criteria. (2) Second, biases (i.e., deviations between true and reestimated parameters) were examined. We were also interested in the influence of the true value of the parameter on the bias, as a positive (negative) relationship can lead to overestimation (underestimation) of a difference between conditions. (3) Third, for the design with two drift rates, we additionally performed power analyses to elicit indications of the number of participants required for the detection of a drift rate difference. (4) The precision of estimation was our fourth criterion. Recently, the idea of using diffusion model parameters for individual diagnostics has been introduced (Aschenbrenner et al., 2016; Ratcliff & Childers, 2015). Certainly, with such an aim it is crucial that parameters be estimated precisely for each person. As a measure of precision, we computed squared deviations of the recovered parameter values from the true values. Thereby—in contrast to the bias measure—over- and underestimations would not cancel each other out. In addition, each parameter's squared deviation was standardized, thereby taking into account the different scales of the different parameters. As a standard value for each parameter, the best-possible accuracy was used, which was defined from an optimal condition of parameter recovery (5,000 trials, at least 4 % of trials at each threshold, no contaminants, and using ML for parameter recovery). On the basis of this measure of parameter recovery, we propose guidelines for how many trials are required for low or high precision in parameter recovery.

Criterion 1: Correlations between true and reestimated parameter values

Regarding the correlations between true and reestimated parameters, all methods turned in a satisfying performance, with the exception of CS performing worse in small samples.

Criterion 2: Parameter estimation biases

In terms of biases, it is noteworthy that biases sometimes decrease with the number of trials. In contrast, for the three- and four-parameter models with fast contaminants, ML and HDDM showed increasing overestimation of the threshold separation and increasing underestimation of nondecision times. This pattern was not observed for the more complex models. We suppose that the intertrial variability of the nondecision time (present in both the six- and seven-parameter models) helped to capture the negative effects of fast contaminants. Note that the decreasing and increasing biases are in contradiction with the hypothesis that only the standard deviation, but not the bias, changes with trial numbers (van Ravenzwaaij & Oberauer, 2009). Importantly, the biases get stable at around 200 to 500 trials. Thus, a further increase in trial numbers does not have a notable influence on the size of the bias.

The trial numbers also sometimes had an influence on the relationship between the true parameter value and the bias. For example, for data with slow contaminants, the relationship between the true value of the threshold separation and the bias increased notably with the number of trials (e.g., from $r = –.07$ for $n = 24$ up to $r = .89$ for $n = 5{,}000$ for ML estimation in the three-parameter model). Where positive relationships between true parameter values and bias lead to an overestimation of the true effect, negative relationships make it more difficult to detect a true difference in parameters. The starting point reveals a consistent pattern of such negative relationships. Thus, the detection of a significant difference in z_r between conditions would be impeded. For the nondecision time, on the other hand, there were no relationships between the size of the true value and the bias.

Criterion 3: Numbers of participants required for detection of a drift rate difference

The most important finding in terms of our power analyses is that an increase in trial numbers beyond 500 trials does not lead to essential further reductions in the requisite number of participants. Interestingly, EZ-based model fits proved to have a high power to detect differences between drift rates. For small trial numbers, EZ outperformed KS, CS, and HDDM. Only ML performed better than EZ.

Criterion 4: Estimation precision

On the basis of the squared deviations between the true and reestimated values, we defined criteria for requisite trial numbers. The results reveal that in the absence of contaminants, parameters can be accurately recovered even with small trial numbers. Analyses for the separate parameters showed that the required trial number was lowest for nondecision times, whereas a precise estimation of the threshold separation required especially high trial numbers. For the condition with no contaminants, HDDM usually led to the most precise estimates, followed by ML and KS. CS showed the worst results. Again, for the three-parameter model EZ could recover especially the drift rates very precisely. In the three- and six-parameter models, due to the fixed starting point, parameters were estimated also for data sets with fewer than 4 % of trials at one of the two thresholds. However, in this case more trials were required to achieve the same precision.

We now turn to the question of the precision of parameter estimation in the presence of contaminants. When

contaminants are slow, both ML and HDDM still provide better results than the other criteria. With fast contaminants, however, KS outperforms the other criteria in almost all conditions. In particular, ML and HDDM are generally affected strongly by fast contaminants. Even our criterion for low precision was in many conditions never reached—that is, even very high trial numbers could not compensate for the presence of fast contaminants. Interestingly, similar to KS, EZ was barely affected by fast contaminants.

We also investigated up to which point additional trials appreciably increase the accuracy of the results. As the slope of the relationship of trial number on precision decreases, increasing the trial number becomes less and less advantageous. Therefore, exceeding a certain number of trials is of limited utility, because the costs will probably be greater than the benefits. For example, it is plausible for the number and percentage of contaminants to increase when participants get tired or bored in long experimental sessions. Splitting sessions over a number of days may also cause problems, since performance may vary from one day to another depending on fatigue, motivation, mood, and so forth. A slope of –0.01 was used to define the point at which more trials did not increase precision notably. Most importantly, the results revealed that it is usually not advisable to increase the number of trials to many hundreds or even thousands, as this improves parameter recovery only marginally.

Number of parameters

The results of our study also provide some insights into the role of additional free parameters. From the three- to the seven-parameter models, there was mostly an increase in the trial numbers required. These results are in line with our finding that the inclusion of a parameter modeling the proportion of contaminants did not lead to any consistent improvement in parameter recovery. In the comparison of the design with one drift rate to the design with two drift rates the additional parameter had a negative effect on threshold separation and drift rate. However, nondecision time was estimated very similarly in both designs, and the starting point was estimated even better in the two-drift design.

One topic urgently calling for further exploration is the poor estimation of the intertrial variabilities s_{zr} and s_{ν}. Even in the condition with 5,000 trials, parameter estimates of s_{zr} and s_{ν} displayed correlations with true values lower than .50 (for similar results, see Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007; van Ravenzwaaij & Oberauer, 2009). Typically, these parameters are included in the model to explain fast (s_{zr}) or slow (s_{ν}) error RTs. One study in which the role of the intertrial variabilities has been explicitly tackled was conducted by Lerche and Voss (manuscript under review). They examined the question of whether fixing these parameters at zero might result in better overall estimation of the remaining parameters, even if there is moderate variability in the true parameter values. To this end, they compared differently complex parameter models analyzing both simulated and empirical data sets. The results showed that the seven-parameter model often provides poorer results than less complex models. In line with these findings is a study by van Ravenzwaaij, Donkin, and Vandekerckhove (manuscript under review), who compared the power to detect parameter differences between EZ (Wagenmakers et al., 2007) and a full diffusion model estimation (i.e., inclusive of all three intertrial variabilities). Although the data-generating model included intertrial variabilities, the EZ model (ignoring these variabilities) led to better power than the more complex model for the detection of differences in drift rate and threshold separation. Note that in our analyses, EZ also proved to be very good at estimating drift rates.

Choice of estimation procedure

It is important to note that our results cannot provide one clear-cut answer to the questions of which estimation method should be used and how many trials are required. Several aspects (e.g., type of contamination, presence of intertrial variabilities) have an influence on which method will produce the most reliable results. In the following, we shortly sketch some guidelines that can help researchers to make qualified decisions for their analyses.

First, researchers have to think about an appropriate experimental paradigm to analyze their research question. Several experimental paradigms have already been analyzed in terms of validity (experimentally or by means of correlations with external criteria). Completely new paradigms should first be validated before applying them to analyze new research questions. Note that in our study we only analyzed two rather simple experimental designs (one-drift and two-drift designs). We suppose, however, that the main patterns of results will remain similar (e.g., best performance of HDDM/ML for uncontaminated data and of KS in the presence of fast contaminants).

Second, the number of trials of an experiment has to be defined. This question will often be related to the chosen paradigm. Especially, if material is restricted, it might be difficult to compose high trial numbers. The homogeneity of the material also influences the decision process, with more heterogeneous material resulting in higher intertrial variability of the drift. Besides, the researcher has to consider the fatigue that the type of task might cause. For tasks that are very demanding and that take very long, a higher percentage of contaminants is to be expected.

Third, after collecting the data, the researcher should analyze their quality before applying a diffusion model. This means, for example, figuring out whether there are supposedly many contaminants. If, for example, the RT distributions

include many statistical outliers according to typical outlier detection procedures (e.g., Tukey, 1977), this might indicate a high level of contaminants. Note that the exclusion of outliers does not necessarily lead to better parameter estimates, as our additional analyses showed. The problem is that not all contaminants will be detected (especially not if they are situated overlapping with the true RT distribution), and "real" RTs might be accidentally removed from the distribution (false positives). Thus, an estimation method that is robust to contaminants (like KS) is in such cases more adequate than an overly strict data cleaning. Besides, estimation of the intertrial variability of the nondecision time (but not of the rather poorly estimated other two variability parameters) can help to counteract the influence of fast contaminants.

Furthermore, one can analyze whether there might be a response bias for one of the two stimuli. If, for example, correct responses to stimulus A are faster than errors, whereas for stimulus B the errors are faster than the correct responses, the starting point might be positioned closer to stimulus A than to B. In such a case, the researcher should not collapse over the two stimuli by estimation of a model with correct and error responses at the two thresholds. Rather, he or she should use a model with the two different stimuli at the thresholds and freely estimate the starting point. Besides, an analysis of the mean RTs of correct and error responses can give a hint as to whether there might be high intertrial variability in the data. Finally, on the basis of these analyses, the researcher can decide which parameters to estimate and which estimation method to use.

Thus, one main message of this article is that there is no single type of diffusion model analysis. Several aspects influence the parameter estimation, and thus, the estimation procedure has to be carefully selected. Our work is intended as a first step in the development of general guidelines for diffusion modeling.

Conclusions

Whereas several hundred or even several thousand trials are often used in the application of the Ratcliff diffusion model (Ratcliff, 1978), our simulation studies—executed with the newest version of fast-dm (Voss et al., 2015)—indicate that in most cases considerably lower trial numbers are sufficient. Using a lot more than the necessary number of trials can also be more detrimental than useful. It leads to higher costs (e.g., longer preparation and execution time of the experiment, or fatigue of the participants) without clearly improving parameter estimation performance. In this article, we give guidelines for the number of trials required, depending on the optimization criterion applied, the number of parameters estimated, and the presence of contaminants.

Our simulations provide the following stable patterns of results: (1) CS is generally not advisable for small to moderate trial numbers; (2) parameter recovery often does not improve much if more than around 500 trials are used; (3) for less complex models (i.e., exclusive of intertrial variabilities), notably smaller trial numbers are sufficient; (4) ML and HDDM perform best for uncontaminated data; and (5) KS and EZ are the methods least affected by fast contaminants.

Author note This research was supported by a grant from the German Research Foundation to A.V. (Grant No. VO1288/2-1).

References

Arnold, N. R., Bröder, A., & Bayen, U. J. (2015). Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychological Research, 79,* 882–898. doi:10.1007/s00426-014-0608-y

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2016). A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. *Neuropsychology, 30,* 225–238. doi:10.1037/neu0000222

Boywitt, C. D., & Rummel, J. (2012). A diffusion model analysis of task interference effects in prospective memory. *Memory & Cognition, 40,* 70–82. doi:10.3758/s13421-011-0128-6

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178. doi:10.1016/j.cogpsych.2007.12.002

Champely, S. (2012). pwr: Basic functions for power analysis (R package version 1.1.1). Available at http://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16,* 1129–1135. doi:10.3758/PBR.16.6.1129

Fifić, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117,* 309–348. doi:10.1037/a0018526

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed–accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience, 31,* 17242–17249. doi:10.1523/JNEUROSCI.0309-11.2011

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision making: A diffusion model analysis. *Personality and Social Psychology Bulletin, 40,* 217–231.

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology, 53,* 55–68. doi:10.1016/j.jmp.2009.01.006

Heathcote, A., & Brown, S. (2004). Reply to Speckman and Rouder: A theoretical basis for QML. *Psychonomic Bulletin & Review, 11,* 577–578.

Horn, S. S., Bayen, U. J., & Smith, R. E. (2011). What can the diffusion model tell us about prospective memory? *Canadian Journal of Experimental Psychology, 65,* 69–75. doi:10.1037/a0022808

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology, 93,* 353–368. doi:10.1037/0022-3514.93.3.353

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Oxford: Academic Press.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6,* 651–687.

Lerche, V., & Voss, A. (in press). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*. doi:10.1007/s00426-016-0770-5

McKoon, G., & Ratcliff, R. (2013). Aging and predicting inferences: A diffusion model analysis. *Journal of Memory and Language, 68,* 240–254. doi:10.1016/j.jml.2012.11.002

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology, 27,* 193–200. doi:10.1037/a0031533

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion, 13,* 739–747. doi:10.1037/a0031628

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. **Retrieved from www.R-project.org**

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. doi:10.1037/0033-295x.85.2.59

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychonomic Bulletin & Review, 15,* 1218–1228. doi:10.3758/PBR.15.6.1218

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2,* 237–279. doi:10.1037/dec0000030

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. doi:10.1111/1467-9280.00067

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481. doi:10.3758/bf03196302

Ratcliff, R., & Van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin & Review, 16,* 742–751. doi:10.3758/PBR.16.4.742

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19,* 278–289. doi:10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50,* 408–424. doi:10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60,* 127–157. doi:10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140,* 464–487. doi:10.1037/a0023810

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 222–250. doi:10.1037/a0026003

Schmitz, F., & Voss, A. (2014). Components of task switching: A closer look at task switching and cue switching. *Acta Psychologica, 151,* 184–196. doi:10.1016/j.actpsy.2014.06.009

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence, 51,* 28–46. doi:10.1016/j.intell.2015.05.002

Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 101–117. doi:10.1037/0278-7393.32.1.101

Speckman, P. L., & Rouder, J. N. (2004). A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. *Psychonomic Bulletin & Review, 11,* 574–576.

Swensson, R. G. (1972). The elusive tradeoff: Speed vs. accuracy in visual discrimination tasks. *Perception & Psychophysics, 12,* 16–32. doi:10.3758/bf03212837

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53,* 463–473. doi:10.1016/j.jmp.2009.09.004

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14,* 1011–1026. doi:10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40,* 61–72. doi:10.3758/BRM.40.1.61

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16,* 44–62. doi:10.1037/a0021765

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767–775. doi:10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52,* 1–9. doi:10.1016/j.jmp.2007.09.005

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32,* 1206–1220. doi:10.3758/BF03196893

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology, 44,* 1048–1056. doi:10.1016/j.jesp.2007.10.009

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63,* 539–555. doi:10.1348/000711009x477581

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60,* 385–402. doi:10.1027/1618-3169/a000218

Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in associative and categorical priming: A diffusion model analysis. *Journal of Experimental Psychology: General, 142,* 536–559. doi:10.1037/a0029459

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology, 6,* 336. doi:10.3389/fpsyg.2015.00336

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods, 46,* 15–28. doi:10.3758/s13428-013-0369-3

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14,* 3–22. doi:10.3758/bf03194023

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision

task. *Journal of Memory and Language, 58,* 140–159. doi:10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review, 15,* 1229–1235. doi:10.3758/PBR.15.6.1229

White, C. N., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion-model analysis. *Cognition and Emotion, 23,* 181–205. doi:10.1080/02699930801976770

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010a). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion, 10,* 662–677. doi:10.1037/a0019474

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010b). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54,* 39–52. doi:10.1016/j.jmp.2010.01.004

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics, 7,* 14. doi:10.3389/fninf.2013.00014

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 53–79. doi:10.1037/a0024177

Yap, M. J., Balota, D. A., & Tan, S. E. (2013). Additive and interactive effects in semantic priming: Isolating lexical and decision processes in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 140–158. doi:10.1037/a0028520